Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society

1 – 4 August 2001

University of Edinburgh
Edinburgh, Scotland

Editors:
Johanna D. Moore
Keith Stenning

Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society

Johanna D. Moore and Keith Stenning
Editors

August 1-4, 2001
Human Communication Research Centre
University of Edinburgh
Edinburgh, Scotland

LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
2001 Mahwah, New Jersey London

Distributed by
Lawrence Erlbaum Associates, Inc.
10 Industrial Avenue
Mahwah, New Jersey 07430

ISBN 0-8058-4152-0

ISSN 1047-1316

Printed in the United States of America

Dedicated to the memory of Herbert A. Simon,
June 15, 1916 – February 9, 2001

How It All Got Put Together

Once upon a time when the world
was young,
Oh best beloved.
There came to the banks of the Monongogo River,
All muddy and brown,
Oh best beloved,
A djinn who was one thing on the inside
But many things on the outside.
And he camped by the banks of the Monongogo River,
All muddy and brown,
Oh best beloved.
And he stayed and stayed and he never went away.
And he did his magic there.
He had many hands, each hand with many fingers,
Oh best beloved.
More hands and fingers than you and I
More hands than you have fingers,
More fingers on each hand than you have toes.
Each hand played a tune on a magic flute,
Oh best beloved.
And each fluted tune floated out on a separate flight.
And each was a tune for a separate dance,
And each was heard in a separate place,
And each was heard in a separate way,
And each was merged in the dance it swayed.
But it was still all the same tune,
For that was the magic of the djinn.
Now, best beloved, listen near—
Each separate place, when the world was young,
Danced in a way that was all its own,
Different from all of the others.
But the melody told of how it could be
That creatures out of an ancient sea,
By dancing one dance on the inside,
Could dance their own dance on the outside,
Because of the place where they were in—
All of its ins and outs.
For that was the magic of the djinn.
And little by little, each swayed a new way,
Taking the melody each its own way,
But hearing the melodies far away
From other places with separate dances,
But the very same melody
That told the dance to be done on the inside.
So, each started to step in the very same way,
Putting together one dance on the inside
For many dances on the outside.
So the melody grew, and it drifted back
To the Monongogo River, all muddy and brown,
And the river came clear and sweet.
Ah, best beloved, I must tell the truth.
The river is not yet clear and sweet,
Not really so.
Because putting together is a task forever.
And no one—not even a djinn with kilohands and megafingers,
All of which play a different-same tune—
Can put all things together in a single breath,
Not even a breath of fifty years.
It is not all put together yet,
And it never shall be,
For that is the way of the world.
But even so, when the world was young,
Was the time of the need for the single tune
To guide the dance that would move together
All of the steps in all of the places.
And it happened by the banks of the Monongogo River,
All muddy and brown,
Best beloved.
And the river will never be the same.
Just so.

Allen Newell
Carnegie Mellon University

Foreword

This volume contains the papers and posters selected for presentation at the 23rd Annual Meeting of the Cognitive Science Society in Edinburgh, August 1–4th 2001. This meeting is the first in the history of the society to be held outside North America, and reflects the increasing internationalisation of cognitive science. More than 500 submissions were received from all over the world. The breadth of topics treated, together with the evident themes that recur are testimony to the development of a distinctive field. We were reminded of the multidimensionality of the field when the several papers on topics related to categorisation proved to be the hardest of all to categorise.

It is our belief that the virtue of cognitive science comes from its deep engagement with the full range of disciplines that contribute to informational theories of mind. Cognitive science began with the realisation that several disciplines studied what is ultimately the same subject matter using different concepts and methods. Observation and experiment had become separated from simulation, engineering, formal analysis, historical, cultural and evolutionary study, and philsophical speculation. It is our hope that this conference will play its small part in substantiating the vision that it is important to put back together what the disciplines have cast asunder.

This multidimensionality of the field makes scheduling a major headache. It is impossible to ensure that clashes do not occur. At one point in scheduling we resorted to statistical corpus analysis on the presented papers to reveal implicit structure. (You will perhaps be relieved to hear that human analysis still appears to be ahead of LSA at this task). We hope that you enjoy the program that has resulted.

We would like to acknowledge help from the following sources, without whom this event would certainly not have been possible:

The Cognitive Science Society Board for inviting us to host the meeting and providing the framework, expertise and support.

The Program Committee assigned submissions to referees, read their resulting reviews and made judgments on the five hundred submissions.

The Reviewers (and there were more than 250 of them) reviewed the papers and gave feedback to committee and authors. Interdisciplinary reviewing is not an easy task. Submitting interdisciplinary papers sometimes feels like being tried by the legal systems of several cultures simultaneously. A necessarily imperfect process was carried out with good grace and some assurance of the quality of decisions. These tasks of assigning and performing reviews are second only to the quality of submissions in determining the calibre of the meeting.

The Tutorial Chair (Frank Ritter) who was responsible for the construction and organisation of the tutorial program.

The many volunteers who helped with the myriad local arrangements for a meeting of this size, and especially Jean McKendree who chaired the local arrangements committee.

The meeting certainly would not have happened without Frances Swanwick who coordinated the submissions process and Jonathan Kilgour who kept the software working, or without Mary Ellen Foster's tireless work on the Proceedings.

Janet Forbes and her successor David Dougal, and their secretarial team: Margaret Prow, Eva Steel, and Yvonne Corrigan for providing administrative support.

Financial support: British Academy, NSF, Erlbaum, Elsevier, Wellcome, the Glushko Foundation, and the Human Communication Research Centre.

The plenary speakers Jon Elster, Wilfred Hodges and Dan Sperber.

And lastly, and most importantly, the authors and symposium participants who presented their work, and made the conference what it was.

Johanna Moore and Keith Stenning
Conference Chairs, CogSci 2001
Human Communication Research Centre
Edinburgh University

Twenty-Third Annual Conference of the Cognitive Science Society

August 1-4 2001
Human Communication Research Centre
University of Edinburgh
Scotland

Conference Co-Chairs

Johanna D. Moore, University of Edinburgh
Keith Stenning, University of Edinburgh

Conference Program Committee

Susan Brennan, SUNY Stonybrook
Gordon Brown, Warwick
Nick Chater, Warwick
Peter Cheng, Nottingham
Andy Clarke, Sussex
Axel Cleeremans, Brussels
Gary Cottrell, UCSD
Matt Crocker, Saarbrucken
Jean Decety, INSERM, Paris
Rogers Hall, UC Berkeley
Dan Jurafsky, U. Colorado, Boulder
Irvin Katz, ETS, Princeton
Ken Koedinger, CMU
Michiel van Lambalgen, Amsterdam
Frank Ritter, Penn State
Mike Oaksford, Cardiff
Stellan Ohlsson, U. Illinois, Chicago
Tom Ormerod, Lancaster
Michael Pazzani, UC Irvine
Christian Schunn, George Mason
Steven Sloman, Brown University
Niels Taatgen, Groningen
Andree Tiberghien, CNRS, Lyon
Richard Young, Hertfordshire
Jiajie Zhang, U. Texas at Houston

Local Arrangments Committee

Jean McKendree, Chair
David Dougal
Janet Forbes
Ian Hughson
Padraic Monaghan
Peter Wiemer-Hastings
Daniel Yarlett

Submissions Coordinator Frances Swanwick
Conference Software Maintainer Jonathan Kilgour
Proceedings Mary Ellen Foster, Jonathan Kilgour
Program Coordinator Michael Ramscar
Registration Website Arthur Markman
Website John Mateer, Jonathan Kilgour, Frances Swanwick

Marr Prize 2001

Sam Scott, Department of Cognitive Science, Carleton University
Metarepresentation in Philosophy and Psychology

This conference was supported by the Cognitive Science Society, The British Academy, The Wellcome Foundation, Lawrence Erlbaum Associates Ltd, Elsevier Science, The Glushko Foundation and The Human Communication Research Center.

The Cognitive Science Society

The Cognitive Science Society, Inc., was founded in 1979 to promote interchange across traditional disciplinary lines among researchers investigating the human mind. The Society sponsors an annual meeting, and publishes the journal Cognitive Science. Membership in the Society requires a doctoral degree in a related discipline (or equivalent research experience); graduate and undergraduate students are eligible for a reduced rate membership; and all are welcome to join the society as affiliate members. For more information, please contact the society office or see their web page at *http://www.cognitivesciencesociety.org/*

Cognitive Science Society, University of Michigan, 525 East University, Ann Arbor MI, 48109-1109; cogsci@umich.edu; phone and fax (734) 429-4286

Reviewers for the Twenty-Third Annual Conference of the Cognitive Science Society

Agnar Aamodt
Amit Almor
Rick Alterman
Eric Altmann
Richard Anderson
Jennifer Arnold
Stephanie August
Neville Austin
Thom Baguley
Todd Bailey
Nicolas Balacheff
Linden Ball
Dale Barr
Pierre Barrouille
Renate Bartsch
Rik Belew
Bettina Berendt
Rens Bod
Lera Boroditsky
Heather Bortfeld
Brian Bowdle
Holly Branigan
Frances Brazier
Bruce Bridgeman
Ted Briscoe
Paul Brna
Andrew Brook
Patty Brooks
Curtis Brown
Marc Brysbaert
John Bullinaria
Curt Burgess
Bruce Burns
Ruth Byrne
Antonio Caballero
Laura Carlson
Mei Chen
Morten Christiansen
Ed Chronicle
James Chumbley
Cathy Clement
Charles Clifton
Tom Conlon
Fred Conrad
Rick Cooper
Richard Cox
Ian Cross
Fernando Cuetos
Matthew Dailey
Helen deHoop
Arnaud Destrebecqz
Morag Donaldson
Ann Dowker
Ben du Boulay
Reinders Duit
George Dunbar
Matthew Elton
Randi Engle
Mary Enright
Noel Enyedy
Michael Erickson
Martha Evens
John Everatt
Neil Fairley
Marte Fallshore
Vic Ferreira
Rodolfo Fiorini
Ilan Fischer
Peter Flach
Nancy Franklin
Robert French
Ann Gallagher
Simon Garrod
Mike Gasser
Richard Gerrig
David Glasspool
Fernand Gobet
Laura Gonnerman
Barbara Gonzalez
Peter Gordon
Barbara Graves
Wayne Gray
Peter Grunwald
Prahlad Gupta
Karl Haberlandt
Constantinos Hadjichristidis
Fritz Hamm
James Hampton
Joy Hanna
Trevor Harley
Cathy Harris
Nancy Hedberg
Neil Heffernan
Evan Heit
Petra Hendriks
Denis Hilton
Eduard Hoenkamp
Ulrich Hoffrage
Douglas Hofstadter
Anne Holzapfel
Sid Horton
Eva Hudlicka
Elizabeth Ince
Heisawn Jeong
Michael Kac
James Kahn
Hans Kamp
David Kaufman
James Kaufman
Fred Keijzer
Frank Keller
Gerard Kempen
Thomas King
Sheldon Klein
Guenther Knoblich
Chris Koch
Derek Koehler
Boicho Kokinov
Rita Kovordanyi
Carol Krumhansl
Pat Kyllonen
Aarre Laakso
Nicki Lambell
Matthew Lambon-Ralph
Donald Laming
Alex Lamont
Peter Lane
Maria Lapata
Michal Lavidor
John Leach
David Leake
Christian Lebiere
Jeff Lidz
Brad Love
Will Lowe
George Luger
Jose Luis Bermudez
Rachel McCloy
Scott McDonald
Brendan McGonigle
Jim MacGregor
Jean McKendree
Craig McKenzie
Brian MacWhinney
Paul Maglio
Lorenzo Magnani
Barbara Malt
Ken Manktelow
Denis Mareschal
Art Markman
Amy Masnick
Santosh Mathan
Yoshiko Matsumoto
Mark Mattson
Sven Mattys
David Medler
Monica Meijsing
Paola Merlo
Craig Miller
Toby Mintz
S Mitra
Naomi Miyake
Padraic Monaghan
Joyce Moore
Bradley Morris
Paul Munro
Wayne Murray
Srini Narayanan

J Nerbonne
David Noelle
Breanndan O Nuallain
Padraig O'Seaghdha
Magda Osman
Helen Pain
Leysia Palen
Barbara Partee
Vimla Patel
Kevin Paterson
Barak Pearlmutter
David Peebles
Pierre Perruchet
Alexander Petrov
Steven Phillips
Massimo Piattelli-Palmarini
Martin Pickering
Julian Pine
Massimo Poesio
Eric Postma
Emmanuel Pothos
Athanassios Protopapas
Michael Ramscar
William Rapaport
Stephen Read
Bob Rehder
Kate Rigby
Steve Ritter
Bethany Rittle-Johnson
Max Roberts
Scott Robertson
Jenni Rodd
Robert Roe
Christoph Scheepers
Hermi Schijif
Friederike Schlaghecken
Matthew Schlesinger
Ute Schmid
Thomas Schultz
Philippe Schyns
Julie Sedivy
David Shanks
Bruce Sherin
Val Shute
Asma Siddiki
Derek Sleeman
Peter Slezak
Vladimir Sloutsky
Linda Smith
Cristina Sorrentino
Jacques Sougne
Bobbie Spellman
Michael Spivey
Constance Steinkuehler
Suzanne Stevenson
Neil Stewart
Stephen Stich
Rob Stufflebeam
Patrick Sturt
Michael Tanenhaus
Heike Tappe
Adam Taylor
Virginia Teller
Josh Tenebaum
Charles Tijus
Michael Tomasello
Greg Trafton
David Traum
Jody Underwood
Ludger van Elst
Ezra van Everbroeck
Maarten van Someren
Alonso Vera
Rineke Verbrugge
Gregg Vesonder
Michael Waldmann
Lyn Walker
William Wallace
Hongbin Wang
Pei Wang
Amy Weinberg
Mike Wheeler
Bob Widner
Cilia Witterman
Amanda Woodward
Lee Wurm
Takashi Yamauchi
Wai Yeap
Wayne Zachary
Jeff Zacks
Corrine Zimmerman
Daniel Zizzo

Tutorial Program

August 1st, 2001

How to Deal with Modularity in Formal Language Theory: An Introduction to Grammar Systems, Grammar Ecosystems and Colonies

Carlos Martin-Vide, Rovira i Virgili University

APEX: An Architecture for Modeling Human Performance in Applied HCI Domains

Michael Matessa, NASA Ames Research Center
Michael Freed - NASA Ames Research Center
John Rehling - NASA Ames Research Center
Roger Remington - NASA Ames Research Center
Alonso Vera - NASA Ames Research Center

An Introduction to the COGENT Cognitive Modelling Environment (with special emphasis on applications in computational linguistics)

Dr. Richard Cooper, Birkbeck College
Dr. Peter Yule, Birkbeck College

Eye Tracking

Roger P.G. van Gompel, University of Dundee
Wayne S. Murray, University of Dundee

ACT-R 5.0

John R. Anderson, Carnegie Mellon University

Tutorial Co-Chairs

Frank Ritter, Penn State University
Richard Young, University of Hertfordshire

Tutorial Committee Members

Randy Jones, University of Michigan
Todd Johnson, University of Texas, Houston
Vasant Honavar Iowa State University
Kevin Korb, Monash University
Michail Lagoudakis, Duke University
Toby Mintz, University of Southern California
Josef Nerb, University of Freiberg and University of Waterloo
Gary Jones, University of Derby
Padraic Monaghan, University of Edinburgh

Speakers and Symposia

Invited Speakers

Jon Elster, Columbia University
Wilfred Hodges, Queen Mary and Westfield College, University of London
Dan Sperber, CNRS, Paris

Invited Symposia

Emotion and Cognition
Chair: Keith Stenning, University of Edinburgh
Speakers:
Ziva Kunda, Waterloo University
Paul Seabright, Toulouse University
Drew Westen, Boston University

Representation and Modularity
Chair: Jon Oberlander, University of Edinburgh
Speakers:
Lawrence Hirschfeld, University of Michigan
Annette Karmiloff-Smith, Institute of Child Health, London
Dylan Evans, King's College, London

Submitted Symposia

Computational Models of Historical Scientific Discoveries
Chairs:
Pat Langley, ISLE, Stanford
Lorenzo Magnani, University of Pavia
Presenters:
Peter Cheng, Adrian Gordon, Sakir Kocabas, Derek Sleeman

When Learning Shapes its own Environment
Chair:
James Hurford, University of Edinburgh
Presenters:
Gerd Gigerenzer, Simon Kirby, Peter Todd

The Interaction of Explicit and Implicit Learning
Chairs:
Ron Sun, University of Missouri-Columbia
Robert Matthews, Louisiana State University
Presenters:
Axel Cleermans, Zoltan Dienes

The Cognitive Basis of Science: The View from Science
Chair:
Nancy Nersessian, Georgia Institute of Technology
Presenters:
Stephen Sich, Ronald Giere, Dedre Gentner

Herb Simon Memorial Symposium

Chair: John Anderson

Presenters:
Pat Langley, ISLE, Stanford
"Computational Scientific Discovery and Human Problem Solving"

Fernand Gobet, University of Nottingham
"Is Experts' Knowledge Modular?"

Kevin Gluck, Air Force Research Laboratory
"The Right Tool for the Job: Information Processing Analysis in Categorisation"

"For us life is, as Shakespeare and many others have described it, a play—a very serious play whose meaning lies in living it. Like any play, in order to have meaning, it must have a beginning, a middle and an end. If an act spans about a decade, eight acts are already a very long play, making heavy demands on the dramatist (oneself) to give it shape.

"Dot and I have had remarkably happy and lucky lives (the first requires the second), which continue to be interesting and challenging, and we have no urge to end them. On the other hand, the realization that these lives are likely, in fact, to end at almost any time now evokes no resentment of fate—at most, sometimes a gentle sadness. We are resigned, not in a sense of giving up or losing, but in a sense of wanting to end our years with dignity, good memories and a feeling that the play had a proper shape and ending, including a final curtain."

Herb Simon

Contents

Symposia

Papers & Posters

Referential Form, Word Duration, and Modeling the Listener in Spoken Dialogue
Ellen Bard and Matthew Aylett (University of Edinburgh)

The Utility of Reversed Transfers in Metaphor
John Barnden (The University of Birmingham)

A model theory of deontic reasoning about social norms
Sieghard Beller (Department of Psychology, University of Freiburg, Germany)

Cue Preference in a Multidimensional Categorization Task
Patricia Berretty (Fordham University)

A Perceptually Driven Dynamical Model of Rhythmic Limb Movement and Bimanual Coordination
Geoffrey Bingham (Psychology Department and Cognitive Science Program, Indiana University)

Inferences About Personal Identity
Sergey Blok, George Newman , Jennifer Behr and Lance Rips (Northwestern University)

Graded lexical activation by pseudowords in cross-modal semantic priming: Spreading of activation, backward priming, or repair?
Jens Bölte (Psychologisches Institut II)

Understanding recognition from the use of visual information
Lizann Bonnar, Philippe Schyns and Frédéric Gosselin (University of Glasgow)

Taxonomic relations and cognitive economy in conceptual organization
Anna Borghi (University of Bologna) and Nicoletta Caramelli (University of Bologna)

The Roles of Body and Mind in Abstract Thought.
Lera Boroditsky (Stanford University), Michael Ramscar (Edinburgh University) and Michael Frank (Stanford University)

The time-course of morphological, phonological and semantic processes in reading Modern Standard Arabic
Sami Boudelaa and William Marslen-Wilson (MRC-CBU)

Reference-point Reasoning and Comparison Asymmetries
Brian Bowdle (Indiana University) and Douglas Medin (Northwestern University)

Deference in Categorisation: Evidence for Essentialism?
Nick Braisby (Open University)

Meaning, Communication and Theory of Mind.
Richard Breheny (RCEAL, University of Cambridge)

The Effects of Reducing Information on a Modified Prisoner's Dilemma Game
Jay Brown and Marsha Lovett (Carnegie Mellon University)

Mice Trap: A New Explanation for Irregular Plurals in Noun-Noun Compounds
Carolyn Buck-Gengler, Lise Menn and Alice Healy (University of Colorado, Boulder)

Simulating the Evolution of Modular Neural Systems
John Bullinaria (University of Birmingham, UK)

The Hot Hand in Basketball: Fallacy or Adaptive Thinking?

Constraints on Linguistic Coreference: Structural vs. Pragmatic Factors
Frank Keller (Computational Linguistics, Saarland University) and Ash Asudeh (Department of Linguistics, Stanford University)

Training for Insight: The Case of the Nine-Dot Problem
Trina Kershaw and Stellan Ohlsson (University of Illinois at Chicago)

Theory-based reasoning in clinical psychologists
Nancy Kim (Yale University) and Woo-kyoung Ahn (Vanderbilt University)

Effect of Exemplar Typicality on Naming Deficits in Aphasia
Swathi Kiran, Cynthia Thompson and Douglas Medin (Northwestern University)

Visual Statistical Learning in Infants
Natasha Kirkham, Jonathan Slemmer and Scott Johnson (Cornell University)

Episode Blending as Result of Analogical Problem Solving
Boicho Kokinov and Neda Zareva-Toncheva (New Bulgarian University)

Dissecting Common Ground: Examining an Instance of Reference Repair
Timothy Koschmann (Southern Illinois University), Curtis LeBaron (University of Colorado at Boulder), Charles Goodwin (UCLA) and Paul Feltovich (Southern Illinois University)

Kinds of kinds: Sources of Category Coherence
Kenneth Kurtz and Dedre Gentner (Northwestern University)

Learning Perceptual Chunks for Problem Decomposition
Peter Lane, Peter Cheng and Fernand Gobet (University of Nottingham)

The Mechanics of Associative Change
Mike Le Pelley and Ian McLaren (Department of Experimental Psychology, Cambridge University)

Representation and Generalisation in Associative Systems
Mike Le Pelley and Ian McLaren (Department of Experimental Psychology, Cambridge University)

Costs of Switching Perspectives in Route and Survey Descriptions
Paul Lee and Barbara Tversky (Stanford University)

A Connectionist Investigation of Linguistic Arguments from the Poverty of the Stimulus: Learning the Unlearnable
John Lewis (McGill University) and Jeff Elman (University of California, San Diego)

Ties That Bind: Reconciling Discrepancies Between Categorization and Naming
Kenneth Livingston, Janet Andrews and Patrick Dwyer (Vassar College)

Effects of multiple sources of information on induction in young children
Yafen Lo (Rice University) and Vladimir Sloutsky (Ohio State University)

Activating verb semantics from the regular and irregular past tense.
Catherine Longworth, Billi Randall, Lorraine Tyler (Centre for Speech and Language, Dept. Exp. Psychology, Cambridge, UK.) and William Marslen-Wilson (MRC Cognition and Brain Sciences Unit, Cambridge, UK)

Member Abstracts

Symposium
Abstracts

Computational Models of Historical Scientific Discoveries

Pat Langley, Institute for the Study of Learning and Expertise
Lorenzo Magnani, Department of Philosophy, University of Pavia
Peter C.-H. Cheng, School of Psychology, University of Nottingham
Adrian Gordon, Department of Computing, University of Northumbria
Sakir Kocabas, Space Engineering Department, Istanbul Technical University
Derek H. Sleeman, Department of Computing Science, University of Aberdeen

The discovery of scientific knowledge is one of the most challenging tasks that confront humans, yet cognitive science has made considerable progress toward explaining this activity in terms of familiar cognitive processes like heuristic search (e.g., Langley et al., 1987). A main research theme relies on selecting historical discoveries from some discipline, identifying data and knowledge available at the time, and implementing a computer program that models the processes that led to the scientists' insights. The literature on computational scientific discovery includes many examples of such studies, but initial work in this tradition had some significant drawbacks, which we address in this symposium.

One such limitation was that early research in law discovery ignored the influence of domain knowledge in guiding search. For example, Gordon et al. (1994) noted that attempts to fit data from solution chemistry in the late 1700s took into account informal qualitative models like polymerization and dissociation. They have developed Hume, a discovery system that draws on such qualitative knowledge to direct its search for numeric laws. Hume utilizes this knowledge not only to rediscover laws found early in the history of solution chemistry, but also to explain, at an abstract level, the origins of other relations that scientists proposed and later rejected.

Early discovery research also downplayed the role of diagrams, which occupy a central place in many aspects of science. For example, Huygens' and Wren's first presentations of momentum conservation took the form of diagrams, suggesting they may have been instrumental in the discovery process. In response, Cheng and Simon (1992) have developed Huygens, a computational model for inductive discovery of this law that uses a psychologically plausible diagrammatic approach. The system replicates the discovery by manipulating geometric diagrams that encode particle collisions and searching for patterns common to those diagrams. The quantitative data given to the system are equivalent to those available at the time of the original discovery.

Another challenge concerns the computational modeling of extended periods in the history of science, rather than isolated events. To this end, Kocabas and Langley (1995) have developed BR4, an account of theory revision in particle physics that checks if the current theory is consistent (explains observed reactions) and complete (forbids unobserved reactions), revises quantum values and posits new particles to maintain consistency, and introduces new properties to maintain completeness. BR-4 models, in abstract terms, major developments in particle physics over two decades, including the proposal of baryon and lepton numbers, postulation of the neutrino, and prediction of numerous reactions. Background knowledge about symmetry and conservation combine with data to constrain the search for an improved theory in a manner consistent with the incremental nature of historical discovery.

We hope this symposium will encourage additional research that extends our ability to model historical scientific discoveries in computational terms.

References

Cheng, P. C.-H. and Simon, H. A. (1992). The right representation for discovery: Finding the conservation of momentum. In *Proceedings of the Ninth International Conference on Machine Learning*, pages 62–71, San Mateo, CA. Morgan Kaufmann.

Gordon, A., Edwards, P., Sleeman, D., and Kodratoff, Y. (1994). Scientific discovery in a space of structural models. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pages 381–386, Atlanta. Lawrence Erlbaum.

Kocabas, S. and Langley, P. (1995). Integration of research tasks for modeling discoveries in particle physics. In *Proceedings of the AAAI Spring Symposium on Systematic Methods of Scientific Discovery*, pages 87–92, Stanford, CA. AAAI Press.

Langley, P., Simon, H. A., Bradshaw, G. L., and Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes*. MIT Press, Cambridge, MA.

Symposium: When Cognition Shapes its Own Environment

Peter Todd (ptodd@mpib-berlin.mpg.de)
Center for Adaptive Behavior and Cognition,
Max Planck Institute for Human Development, Berlin, Germany.

Simon Kirby (simon@ling.ed.ac.uk) and **James R Hurford (jim@ling.ed.ac.uk)**
Language Evolution and Computation Research Unit,
Department of Theoretical and Applied Linguistics, University of Edinburgh,
40 George Square, Edinburgh, EH8 9LL, UK.

Introduction

Cognitive mechanisms are shaped by their environments, both through evolutionary selection across generations and through learning and development within lifetimes. But by making decisions that guide actions which in turn alter the surrounding world, cognitive mechanisms can also shape their environments in turn. This mutual shaping interaction between cognitive structure and environment structure can even result in coevolution between the two over extended periods of time. In this symposium, we explore how simple decision heuristics can exploit the information structure of the environment to make good decisions, how simple language-learning mechanisms can capitalize on the structure of the "spoken" environment to develop useful grammars, and how both sorts of cognitive mechanisms can actually help build the very environment structure that they rely on to perform well.

Programme

There will be three talks, as follows:

1. Peter Todd, "Simple Heuristics that exploit environment structure",

   Traditional views of rational decision making assume that individuals gather, evaluate, and combine all the available evidence to come up with the best choice possible. But given that human and animal minds are designed to work in environments where information is often costly and difficult to obtain, we should instead expect many decisions to be made with simple "fast and frugal" heuristics that limit information use. In our study of ecological rationality, we have been exploring just how well such simple decision-making heuristics can do when they are able to exploit the structure of information in specific environments. This talk will outline the research program pursued by the Center for Adaptive Behavior and Cognition as developed in the book, *Simple Heuristics That Make Us Smart* (Oxford, 1999), and highlight how the match between cognitive mechanism structure and environment structure allows the Recognition heuristic and Take The Best heuristic to perform on par with traditionally rational decision mechanisms.

2. Simon Kirby, "The Iterated Learning Model of Language Evolution",

   The past decade has seen a shift in the focus of research on language evolution away from approaches that rely solely on natural selection as an explanatory mechanism. Instead, there has been a growing appreciation of languages (as opposed to the language acquisition device) as complex adaptive systems in their own right. In this talk we will present an approach that explores the relationship between biologically given language learning biases and the cultural evolution of language. We introduce a computationally implemented model of the transmission of linguistic behaviour over time: the Iterated Learning Model (ILM). In this model there is no biological evolution, natural selection, nor any measurement of the success of communication. Nonetheless, there is significant evolution. We show that fully syntactic languages emerge from primitive communication systems in the ILM under two conditions specific to Hominids: (i) a complex meaning space structure, and (ii) the poverty of the stimulus.

3. Peter Todd, Simon Kirby and Jim Hurford, "Putting the Models Together: how the environment is shaped by the action of the recognition heuristic",

   To explore how cognitive mechanisms can exert a shaping force on their environment and thus affect their own performance, we begin by considering the actions of a very simple cognitive mechanism, the recognition heuristic for making choices. This heuristic specifies that when choosing between two options, one of which is recognized and one not, the recognized option should be selected. The recognition heuristic makes good choices, in environments where recognition is correlated with the choice criterion. Many natural environments have this structure, but such structure can also be "built": By using the recognition heuristic, agents can create an environment in which some objects are much more often and "talked about" and recognized than others. An agent-based simulation is used to show what behavioral factors affect the emergence of this environmental structure.

The Cognitive Basis of Science: The View from Science

Session Organizer: **Nancy J. Nersessian (nancyn@cc.gatech.edu)**
College of Computing, 801 Atlantic Drive
Atlanta, GA 30332 USA

The issue of the nature of the processes or "mechanisms" that underlie scientific cognition is a fundamental problem for cognitive science. A rich and nuanced understanding of scientific knowledge and practice must take into account how human cognitive abilities and limitations afford and constrain the practices and products of the scientific enterprise. Reflexively, investigating scientific cognition opens the possibility that aspects of cognition previously not observed or considered will emerge and require enriching or even altering significantly current understandings of cognitive processes.

The Baby in the Lab Coat: Why child development is an inadequate model for understanding the development of science

Stephen P. Stich, Department of Philosophy, Rutgers University

In two recent books and a number of articles, Alison Gopnik and her collaborators have proposed a bold and intriguing hypothesis about the relationship between scientific cognition and cognitive development in childhood. According to this view, the processes underlying cognitive development infants and children and the processes underlying scientific cognition are identical. One of the attractions of the hypothesis is that, if it is correct, it will unify two fields of investigation – the study of early cognitive development and the study of scientific cognition – that have hitherto been thought quite distinct, with the result that advances in either domain will further our understanding of the other. In this talk we argue that Gopnik's bold hypothesis is untenable. More specifically, we will argue that if Gopnik and her collaborators are right about cognitive development in early childhood then they are wrong about science. The minds of normal adults and of older children, we will argue, are more complex than the minds of young children, as Gopnik portrays them. And some of the mechanisms that play no role in Gopnik's account of cognitive development in childhood play an essential role in scientific cognition.

Scientific Cognition as Distributed Cognition

Ronald N. Giere, Center for Philosophy of Science, University of Minnesota

I argue that most important cases of cognition in contemporary science are best understood as examples of distributed cognition. Here I focus exclusively on the acquisition of new knowledge as the paradigm of scientific cognition. Scientific cognition, then, does not reduce to mere distributed computation. The simplest case is that in which two people cooperate in acquiring some knowledge that is not directly acquired by either one alone. It is even possible that neither person could physically perform the task alone. This is an example of what has been called "socially shared cognition" (Resnick) or "collective cognition" (Knorr). The most elaborate example is the case of experimental high-energy physics at CERN, as described by the sociologist, Karin Knorr in her recent book, *Epistemic Cultures*. I go beyond Knorr's analysis to include the particle accelerator and related equipment as part of a distributed cognitive system. So here the cognition is distributed both among both people and artifacts. Such artifacts as diagrams and graphics and even abstract mathematical constructions are also included as components of distributed cognitive systems. This makes it possible to understand the increasing power of science since the seventeenth century as in large measure due to the creation of increasing powerful cognitive systems, both instrumental and representational.

The Cognitive Basis of Model-based Reasoning in Science

Nancy J. Nersessian, Program in Cognitive Science, Georgia Institute of Technology

Although scientific practice is inherently "socially shared cognition," the nature of individual cognitive abilities and how these constrain and facilitate practices still needs to be figured into the account of scientific cognition. This presentation will focus on the issue of the cognitive basis of the model-based reasoning practices employed in creative reasoning leading to conceptual change across the sciences. I will first locate the analysis of model-based reasoning within the mental modeling framework in cognitive science and then discuss the roles of analogy, visual representation, and thought experimenting in constructing new conceptual structures. A brief indication of the lines along which a fuller account of how the cognitive, social, and material are fused in the scientist's representations of the world will be developed. That the account needs to be rooted in the interplay between the individual and the communal in the model-based reasoning that takes place in concept formation and change. Modeling is a principal means through which a scientist transports conceptual resources drawn from her wider cultural milieu into science and transmits novel representations through her community. Scientific modeling always takes place in a material environment that includes the natural world, socio-cultural artifacts (stemming from both outside of science and within it), and instruments devised by scientists and communities to probe and represent that world.

Symposium Discussant: Dedre Gentner, Department of Psychology, Northwestern

The Interaction of Explicit and Implicit Learning

Ron Sun (rsun@cecs.missouri.edu)
University of Missouri-Columbia
Columbia, MO 65203

Robert Mathews (psmath@unix1.sncc.lsu.edu)
Louisiana State University, Baton Rouge
Baton Rouge, LA

The Focus of the Symposium

The role of implicit learning in skill acquisition and the distinction between implicit and explicit learning have been widely recognized in recent years (see, e.g., Reber 1989, Stanley et al 1989, Willingham et al 1989, Anderson 1993), Although implicit learning has been actively investigated, the complex and multifaceted interaction between the implicit and the explicit and the importance of this interaction have not been universally recognized; to a large extent, such interaction has been downplayed or ignored, with only a few notable exceptions. [1] Research has been focused on showing the *lack* of explicit learning in various learning settings (see especially Lewicki et al 1987) and on the controversies stemming from such claims. Similar oversight is also evident in computational simulation models of implicit learning (with few exceptions such as Cleeremans 1994 and Sun et al 2000).

Despite the lack of studies of interaction, it has been gaining recognition that it is difficult, if not impossible, to find a situation in which only one type of learning is engaged (Reber 1989, Seger 1994, but see Lewicki et al 1987). Our review of existing data has indicated that, while one can manipulate conditions to emphasize one or the other type, in most situations, both types of learning are involved, with varying amounts of contributions from each (see, e.g., Sun et al 2000; see also Stanley et al 1989, Willingham et al 1989).

Likewise, in the development of cognitive architectures (e.g., Rosenbloom et al 1993, Anderson 1993), the distinction between procedural and declarative knowledge has been proposed for a long time, and advocated or adopted by many in the field (see especially Anderson 1993). The distinction maps roughly onto the distinction between the explicit and implicit knowledge, because procedural knowledge is generally inaccessible while declarative knowledge is generally accessible and thus explicit. However, in work on cognitive architectures, focus has been almost exclusively on "top-down" models (that is, learning first explicit knowledge and then implicit knowledge on the basis of the former), the bottom-up direction (that is, learning first implicit knowledge and then explicit knowledge, or learning both in parallel) has been largely ignored, paralleling and reflecting the related neglect of the interaction of explicit and implicit processes in the skill learning literature. However, there are a few scattered pieces of work that did demonstrate the parallel development of the two types of knowledge or the extraction of explicit knowledge from implicit knowledge (e.g, Willingham et al 1989, Stanley et al 1989, Sun et al 2000), contrary to usual top-down approaches in developing cognitive architectures.

Many issues arise with regard to the interaction between implicit and explicit processes, which we need to look into if we want to better understand this interaction:

- How can we best capture implicit processes computationally? How can we best capture explicit processes computationally?
- How do the two types of knowledge develop along side each other and influence each other's development?
- Is bottom-up learning (or parallel learning) possible, besides top-down learning? How can they (bottom-up learning, top-down learning, and parallel learning) be realized computationally?
- How do the two types of acquired knowledge interact during skilled performance? What is the impact of that interaction on performance? How do we capture such impact computationally?

Titles of the Talks

Axel Cleeremans: "Behavioral, neural, and computational correlates of implicit and explicit learning"

Zoltan Dienes: "The effect of prior knowledge on implicit learning"

Bob Mathews: "Finding the optimal mix of implicit and explicit learning"

Ron Sun: "The synergy of the implciit and the explicit"

[1] By the explicit, we mean processes involving some form of generalized (or generalizable) knowledge that is consciously accessible.

Papers and Posters

The Roles of Thought and Experience in the Understanding of Spatio-temporal Metaphors

Tracy Packiam Alloway (Tracy.Alloway@ed.ac.uk)
Department of Psychology
University of Edinburgh, 7 George Square Edinburgh EH8 9JZ, UK

Michael Ramscar (michael@cogsci.ed.ac.uk)
School of Cognitive Science, Division of Informatics
University of Edinburgh, 2 Buccleuch Place, Edinburgh, EH8 9LW

Martin Corley (Martin.Corley@ed.ac.uk)
Department of Psychology
University of Edinburgh, 7 George Square Edinburgh EH8 9JZ, UK

Abstract

Spatial and temporal metaphors are often used interchangeably, and thus, offer a unique way of exploring the relationship between language and thought. Both spatial and temporal speaking incorporates two systems of motion. The first is an ego-moving system, when the individual moves from one point to another, spatially, or from the past to the future, temporally. The second is the object- (or time-) moving system, when the individual is stationary and observes objects, or time, moving towards him/her. This study explored the effect of a spatial environment on the ambiguous temporal question: *Next Wednesday's meeting has been moved forward two days--What day is the meeting now?* Results reveal that when participants are immersed in an ego-moving spatial environment, such as a virtual reality game, and receive a prime that causes them to think in an object-moving way, they are more likely to perform a target task in a way consistent with the way they have been primed to think, although it contradicts the spatial motion they subsequently experience in the testing environment.

Introduction

What is the relationship between language and sensory experience? According to one recent claim (Lakoff and Johnson, 1999), abstract concepts, such as time, are substrated in concrete concepts like space that can be experienced directly. The representations of these concrete concepts are formed directly, by experience. Thus, our spatial experiences form "a neural structure that is actually part of, or makes use of, the sensorimotor system of our brains. Much of conceptual inference is, therefore, sensorimotor inference" (Lakoff and Johnson, 1999, p. 20). On this view, our understanding of concepts such as time is predicated on our spatial experiences, and thus the idea of motion in time relies on our understanding of motion in space.

There is evidence for this relationship between motion in space and time in the structure of language. We can talk of *putting things forward* in time, as well as *moving forward* through space (see Lakoff & Johnson, 1999; 1980). According to Lakoff and Johnson's (1980; 1999) Conceptual Metaphor hypothesis, metaphors are not justa manner of speaking but a deeper reflection of human thought processes. Metaphoric speaking is reflective, say Lakoff and Johnson, of deeper conceptual mappings that occur in our thinking and is depicted as an over-arching and general metaphor termed asthe Conceptual Metaphor. Consider the following statements:

Your claims are *indefensible*.

He *attacked every weak point* in my argument.

He *shot down* all of my arguments.

According to the Conceptual Metaphor (metaphoric representation) hypothesis when we use statements such as these we are making use of a larger conglomerate metaphor, in this instance, ARGUMENT IS WAR.[1]

The thrust of the Conceptual Metaphor argument is as follows: arguments are similar to wars in that there are winners and losers, positions are attacked and defended, and one can gain or lose ground. The theory of Conceptual Metaphor suggests that we process metaphors by mapping from a base domain to a target domain. In this particular example, the base domain is ARGUMENT IS WAR and the target domain is asubordinate metaphor such as *Your claims are indefensible.*

Motion in Space and Time

Lakoff and Johnson extend the idea of Conceptual Metaphor to spatio-temporal metaphors by invoking the

[1] Following Lakoff and Johnson's convention (1980), all Conceptual Metaphors are typed in the uppercase to distinguish them from the subordinate metaphors

locative terms of FRONT/BACK to represent how we view time and space. FRONT is assigned on the assumption of motion (Fillmore, 1978). According to this theory, in the *ego-moving* system, FRONT is used to designate a future event because the ego is moving forward and encounters the future event in front of him. In the *time-moving* system, the FRONT term denotes a past event where the ego or the individual is stationary but the events are moving. Thus it is possible to define (at least) two schemas of motion in space.

1) Object-Moving Metaphor (OM)

In this schema of motion,the individual is seen as stationary and objects seem to come towards him/her. For an example of this schema, consider an individual waiting at a bus stop and observing vehicles coming towards him/her. In this schema of motion, the individual assigns the term FRONT to the object closest towards him. In the diagram below, the term FRONT would be assigned to the white rock.

Figure 1

2) Ego-Moving Metaphor (EM)

Inthis schema of motion, the objects are stationary and it is the individual that is in motion. Here, the term FRONT would be assigned to the object furthest away from the individual. In the picture below, it is the black rock that would be labeled as FRONT.

Figure 2

Thus in the EM system, *front* is used to designate an object furthest away from the individual, as the trajectory of motion is in that direction. While in the OM system, the term *front* is assigned to the object closest to the individual.

Motion in Time

The schemas of motion represented in the domain of time reflect the representation of motion in the domain of space.

1) Time Moving metaphor (TM)

The motion of time provides the framework in which temporal metaphors are comprehended. In this schema, *front,* or *ahead* is determined by thefuture moving to the past. For example, in the month of February, Christmas is now in the future. In time it will move to the present and then to the past (e.g. Christmas is *coming*). The individual is a stationary observer as times "flows" past. This schema is the temporal equivalent of the OM metaphor in the domain of space.

2) Ego-Moving metaphor (EM)

The ego or the individual moves from the past to the future such as the sentence His vacation to the beach lay *ahead* of him . In this metaphor, the observer is seen as moving forward through time, passing temporal events that are seen as stationary points. It is thus the temporal equivalent of the spatial EM system, where the observer moves forward through space

When discussing motion in time, temporal events are viewed as points or locations in space, and a similar rationale is used when assigning deictic terms such as *front* and *back*. For example, in the EM system, FRONT is used to designate a future event because the ego is moving forward and encounters the future event in front of him, while in the TM system the FRONT term denotes a past event where the ego or the individual is stationary but the events are moving.

Studies of Spatio-temporal Metaphors

Gentner and Imai (1992), and McGlone and Harding (1998) confirmed the idea that the different schemas of motion (EM and TM in the domain of time) are indeed psychologically real systems.Gentner and Imai found that participants responded faster to questions that were schema consistent with regards to temporal schemas in priming than to questions that were inconsistent with their primes. Gentner and Imai argue that this supports the theory that metaphors are mapped in distinct schemas: the shift from one schema to another causes a disruption in the processing, reflected in increased processing time. They argue that their study indicates that the relations between space and time are reflective of a psychologically real conceptual system as opposed to an etymological relic.[2]

A study by McGlone and Harding (1998) involved participants answering questions about days of the week - relative to Wednesday - which were posed in either the *ego-moving* or the *time-moving* metaphor. *Ego-moving* metaphor trials comprised statements such as We passed the deadline two days ago , whilst *time-moving* metaphor trials involved statements such as The deadline passed us two days ago ; in each case,

[2] Although McGlone and Harding (1998) criticised some aspects of Gentner and Imai s methodology, their corrected replication of the original study confirms its findings.

participants read the statements and were then asked to indicate the day of the week that a given event had occurred or was going to occur. At the end of each block of such priming statements, participants read an ambiguous statement, such as The reception scheduled for next Wednesday has been moved forward two days [3] and then were asked to indicate the day of the week on which this event was now going to occur. Participants who had answered blocks of priming questions about statements phrased in a way consistent with the*ego-moving* metaphor tended to disambiguate moved forward in a manner consistent with the *ego-moving* system (they assigned forward - the front - to the future, and hence thought the meeting had been re-scheduled for Friday), whereas participants who had answered blocks of questions about statements phrased a way consistent with the *time-moving* metaphor tended to disambiguate moved forward in a manner consistent with the *time-moving* system (they assigned forward - the front - to the past, and hence thought the meeting had been re-scheduled for Monday).

This work has been further developed in a recent set of experiments by Boroditsky (2000) which explicitly explored the relationship between the domains of space and time. Boroditsky found that temporal priming significantly influenced temporal reasoning in a cross-domain extension of the paradigm used in earlier experiments. Spatially priming participants with the ego moving schema led them to infer that an ambiguous meeting ("Next Wednesday's meeting has been moved forwards two days") had been moved to Friday, whereas spatially priming participants with the object moving schema led them to assign the meeting to Monday. This study provides good evidence to support the notion that our representation of motion in space is mapped on to our understanding of motion in time, although it leaves open the question of what is directing this representational mapping spatial representations that are contiguous with our embodied experience, or functionally separable, abstract conceptual representations of space and time.

Experiment 1

This experiment directly explores the claim that our embodied experiences in space direct our conceptual understanding of time. Participants were immersed in an embodied environment, a virtual reality game, and were presented with an ambiguous spatial task, either after either a purely embodied prime, or after embodied priming during which a linguistic prime had cued them to *think* in terms of a contrary spatial schema. The experiment was designed to explore the role of experience and thought between the two schemas of motion in the domain of space.

[3] All trials were conducted on a Wednesday.

Participants

61 University of Edinburgh students volunteered to take part in this experiment.

Materials

In order to create a particularly convincing Ego Moving environment, participants played a slightly modified version of a pre-existing section of the virtual reality computer game, *UnReal*. This is a first person perspective game and involves the participant walking through a courtyard environment to complete a task. All monsters and other artifacts of the game that were not relevant to the experiment were removed from this section of the game. The objects in the target task appeared upon completion of the game. These were two chests, with no discernible front or back (unlike other objects, such as a car, or a TV), one of which was closer to the player than the other. The game was projected onto a 368cm by 282cm size screen in order to magnify the virtual effects of the game.

Procedure

Pre-Test

25 participants were tested individually seated in front of the projector screen. The game was set at the point in front of the two chests. Participants did not play the game and were only instructed to *Move to the front chest* . In this condition, the target task was performed in isolation, and the results provided a baseline for how the term *front* in this task is interpreted.

Out of the twenty-five subjects, twelve of them interpreted the term *front* to refer to the chest closest to them, while the rest assigned front to the chest furthest from them, confirming the ambiguity of the assignment of *front* in the target task.

Experimental Conditions

36 participants were tested individually. They were asked to fill in a brief questionnaire requesting demographic information, as well as familiarity with video games and computers. At the end of the questionnaire were the following instructions: Your task is to find the location of a young woman. Try your best to navigate around the environment in order to find her. During this game, it is important to try to remember some key landmarks, such as a **pair of brightly coloured pillars as you enter a path,** as well as the **doors on the buildings.** After you have been playing for some time, you will hear a question requiring a true or false answer. This question will be about the game. Try to answer it correctly and speak your answer loudly."

The participants were then shown how to use the arrow keys on the keyboard when navigating through the environment and then left alone to play the game. (The experimenter was on hand, should the volunteers

have any difficulty maneuvering around the environment; however, all volunteers seemed adequately proficient at navigating around the environment.)

There were two experimental conditions. In the first condition, volunteers received a pre-recorded true/false question specific to the assignment of the term *front* approximately four minutes into playing the game. The question they were posed -- During the game, the green pillar is in front of the red pillar — prompted them to think in an Object Moving manner about space (the green pillar was closer to the participants than the red pillar in the game environment, thus this question is true from an OM perspective).[4] We were interested to see if the thinking in an OM way in answering the question would result in a different assignment of *front* from the EM perspective that was embodied in the game.

The order of the pillars in the question was reversed for half of the participants to counter-act an affirmative response bias. Thus, half the participants answered the true/false question: During the game, the red pillar is in front of the green pillar . The answer to this question was false from an OM perspective.

In the second condition, volunteers received a pre-recorded non-spatial question rather than a spatial prime approximately four minutes into playing the game. They had to provide a true or false answer to the following question: During the game, most of the doors are open". The correct answer to this question was true, however, the amount of doors the volunteer saw depended on the route he or she chose in navigating around the environment to complete the task. However, the question was also presented in the inverse to avoid any particular response bias, and half of the participants in this condition answered the following question: During the game, most of the doors are closed". The question in this condition served as a control to ensure that simply answering a question would not cause people to re-represent their perspective of front or back (but that rather a question must cause people to specifically think in a way that involves a representation of front/back for this to occur).

Playing the EM game served as the embodied prime in this condition.

Once the participants had completed the task, the virtual young woman they sought congratulated them and they were asked to complete the target task: "*Move to the front chest*". The two chests were located on the left of the virtual woman and were added from the *UnReal* directory of furniture to maintain continuity in the environment. Upon completion of the target task, participants were given a short debriefing.

[4] Pre-testing had shown that this question, which is true from an OM perspective, was unambiguous and ordinarily answered from the OM perspective. Out of 20 participants, 90% allocated the term *front* in an OM perspective. A binomial test confirmed this as significant; $p<.001$.

Results

Participants responses for the target task are shown in Figure 3. Out of the total 36 participants, two (5%) did not answer the prime question consistently (i.e., to the OM prime: During the game, the green pillar is in front of the red pillar , they answered false when the correct answer was true). Their data were not used in the following analyses.

Analysis showed that when participants received the OM prime, requiring them to specifically think in a way that represented a particular schema of motion, 75% of them interpreted the *front* chest in an OM consistent manner, despite playing the EM game for a further 2-3 minutes *after* answering the prime question. However, when participants were simply immersed in a game which embodied EM motion, and were not required to specifically (or explicitly) think about that motion (instead, they were required to think about doors), 83%of them were influenced by the nature of motion in the game and interpreted the *front* chest command in an EM schema consistent manner.

A chi-square analysis revealed a significant effect of the type of prime participants received on how they interpreted the term *front* to apply to an ambiguous target task: $\chi^2(1)=11.691$; $p<0.001$.

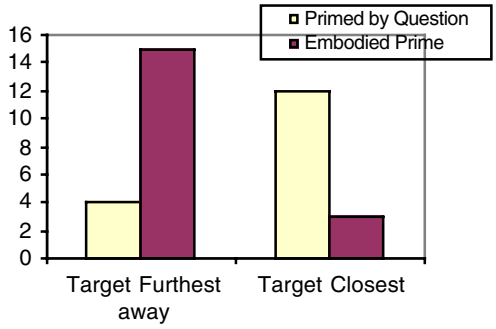


Figure 3: Target responses in each prime condition

Discussion

This experiment seems to suggest that thinking about space can override the role of spatial experience in our understanding of spatial concepts. Participants who were cued to think using a particular schema of motion (OM), overcame the schema of motion they were experiencing (EM), and responded ina consistent manner with the way they had been cued to think about space. Participants who received a random question, unrelated to any system of motion, were influenced by the schema of motion in the game (EM) and responded to the spatial task consistently with their experience.

Although the video prime in this experiment involved the participant in an EM schema of motion, there might be some criticism against the embodied prime as participants only perceived the visual environment, rather than physically experienced it. Such criticism seems unjustified in this case. As Lishman and Lee (1973) argue, perception is powerful enough to direct kinaesthesis or movement. They claim that a person relies heavily on visual kinaesthesis, in many situations, for example, driving, and swimming in a current, a person is *dependent* on vision to sense how he is moving relative to the static environment (p. 288, emphasis theirs). They also argue that individuals experience a sensation of motion when visual scenes change, but they are stationary. They conducted a series of experiments where individuals were placed in a stationary trolley in a moving room. The room was moved independently from the trolley located in it, so the participants saw the room move, although thetrolley they were standing in was completely stationary. Lishman and Lee record participants as perceiving the trolley to move as well, eventhough they were stationary. The participant also swayed together with the room in an apparent attempt to keep himself stable with respect to his static environment (p.292). Participants felt the experience was like being on a boat, and several felt quite nauseated afterwards.

In this virtual reality experiment, several participants had similar experiences and even commented on feeling rather ill after playing the video game for a few minutes. One participant asked if she could leave because she felt so nauseated. Often, participants shoulders were seen to move in sync with a right or left turn they made in the virtual environment, and many participants remarked on feeling dizzy after completing the experiment. This confirms the importance of perception in directing our sense of motion, and suggests that visually experiencing motion in virtual reality provides a similar sensation to a physical experience of motion.

Experiment Two

While the first experiment examined the influence of simple experience versus explicit thought in our understanding of motion in space, this experiment explored whether simple spatial experience or thinking about space would be more influential in mapping information about motion from the domain of space to time. Participants were immersed in an embodied environment and were presented with an ambiguous target temporal task after receiving either a purely embodied priming, or embodied priming during which a linguistic prime had cued them to think in terms of a contrary spatial schema.

Participants

Thirty-nine Edinburgh University students volunteered to take part in this experiment.

Materials

The participants played the same video game as described in Experiment 1.

Procedure

All trials were conducted on a Wednesday. Participants were tested individually in the virtual reality lab and were asked to fill in a brief questionnaire containing the same instructions given in Experiment 1. They were also informed that they would be required to return next Wednesday if they were successful in accomplishing the task in the game. This information provided a connection between the target question (see below) and the experiment, as the participants would interpret Next Wednesday s meeting in the target question as a further experiment, rather than an unrelated question.

Participants were then shown the game and began playing. There were two conditions. Approximately four minutes into playing the game, the participants in the first condition received the linguistic prime cueing them to think the in an OM perspective. Again, they had to respond with either true or false to the question During the game, the green pillar is in front of the red pillar . (Once again, half of the participants in this condition received the inverse question.)

In the second condition, instead of receiving a prime that cued spatial thinking, the participants received the non-spatial question approximately four minutes into playing the game During this game, most of the doors are open . (Again, half of the participants in this condition received the question in the inverse.)

Once participants had successfully completed the game task (finding a virtual young woman; all participants were successful), the experimenter congratulated them and then informed them that "*Next Wednesday's meeting has been moved forward two days* and asked *What day is the meeting now that is has been rescheduled?*" (the ambiguous temporal question used in McGlone and Harding, 1988, and Boroditsky, 2000). After participants had given their answer, they were given a short debriefing.

Results

Out of the total 39 participants, three of the participants (8%) did not answer the prime question consistently (i.e., to the OM prime: During the game, the green pillar is in front of the red pillar , they answered false when the correct answer was true). Data from those participants who provided incorrect answers to the prime were eliminated from the analyses.

Participants responses to the ambiguous target question were examined, and once again, the results revealed that the type of prime participants were presented with significantly affected their disambiguation of the target temporal question. The participants who received the cued OM prime during the game (which required them to adopt an OM schema for thinking of motion in answering the question) were more likely to interpret the term *forward* from Wednesday as Monday (65%) rather than Friday. In comparison, 74% of the participants who were influenced by the embodied EM game, but did not have to explicitly think about schemas of motion in answering the in-game question considered the new meeting day to beFriday rather than Monday (see figure 4).

A chi-square revealed that the type of prime the participants received significantly affected how they disambiguated *forward* in the temporal target task: $\chi^2(1)=5.355$; $p<0.05$ (one-tailed).

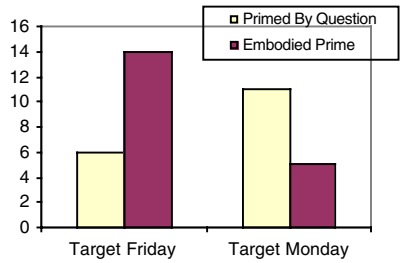


Figure 4: Target responses in each prime condition

Discussion

These results suggest that our concepts of motion in the domain of space can influence how we understand motion in the domain of time. However, while the embodied position suggests that it is our experiences in space that ultimately affect how we think of time, this experiment reveals that how you *think* about motion — even in abstract terms, such as in response to a question — also plays a significant role in influencing our concept of time. Although they participants continued to play a game which embodied an EM spatial perspective *after* they answered their question, participants who answered questions which required them to think in an OM manner answered an ambiguous temporal question in a TM (and thus OM, see Boroditsky, 2000) consistent manner, whereas participants who had played the EM game but not been required to explicitly think about time answered the ambiguous temporal question in an EM consistent manner. This indicates that although spatial experience can influence temporal thought, this influence can be over-ridden by explicitly thinking about space, suggesting that people s conceptual representations of space and time are functionally separable from their embodied experiences of space and time (see also Boroditsky, Ramscar, & Frank, this volume).

General Discussion

In two experiments, we have shown that explicitly thinking about space — in order to provide answers to questions cueing the *object-moving* metaphoric system - could significantly reverse a task bias to assign FORWARD in an *ego-moving* manner.

If, as Lakoff and Johnson (1999) suggest, language is ultimately the slave of our (universal) embodied thought, then we would have expected pure embodied priming to have at least as much an influence as abstract thought. However, this was not the case (see also Boroditsky et al, this volume).

These results suggest that a proper characterization of conceptual thought will need to look beyond the information that comes from physical experience, and consider as well the ways in which languages and cultures affect thought.

Acknowledgements

We thank Lera Boroditsky for many insightful discussions, and Jon Sykes for his assistance in programming the virtual environment.

References

Alloway, T.P., Ramscar, M., & Corley, M. (1999). Verbal and Embodied priming in schema mapping tasks. In *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society.* New Jersey:Lawrence Erlbaum Associates, Pub.

Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition, 75(1)*, 1-28.

Boroditsky, L., Ramscar, M. & Frank, M. (2001). The roles of body and mind in abstract thought. This volume.

Gentner, D., & Imai, M. (1992). Is the future always ahead? Evidence for system mappings in understanding space-time metaphors. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, Bloomington, Indiana, 510-515.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by.* Chicago: Univeristy of Chicago Press.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the Flesh.* New York: Harper Collins Publisher.

McGlone, M., & Harding, J. (1998). Back (or Forward?) to the Future: The Role of Perspective in Temporal Language Comprehension. *Journal of Experimental Psychology, 24,* 1211-1223.

McTaggart, J. (1908). The unreality of time. *Mind, 17*, 457-474.

Coordinating Representations in Computer-Mediated Joint Activities

Richard Alterman, Alex Feinman, Josh Introne, Seth Landsman
Department of Computer Science
Brandeis University
Waltham, MA 02454 USA

Abstract

This paper develops, in the context of the interdisciplinary literature on coordination, the concept of a *coordinating representation* as an everyday method for structuring the coordination of actors engaged in a non face-to-face joint activity. Evidence is provided by applying the idea of coordinating representation to the development of a computer-mediated cooperative activity.

Introduction

A critical reasoning problem confronted by actors as they engage in their everyday activities is the maintenance of coordination (Clark, 1996). Within a community of actors, designs that organize (structure) behavior in recurrent situations of cooperation develop over time. Once developed, the expectation that a given sort of structure might be in place for a given kind of situation simplifies the interaction among the participants while reducing mental effort, physical work, and errors (Alterman & Garland, 1998). In non face-to-face interactions, structures that simplify the coordination of a conventional behavior are coded into a *coordinating representation*. The coordinating representation helps the participants to jointly make sense of the situation in the absence of a face-to-face interaction.

An everyday example of a coordinating representation is the "stop sign". The stop sign is a representation shared among the participants at a traffic setting. The stop sign presents a structure for organizing the collective behavior of drivers, pedestrians, and cyclists at a busy intersection. The interpretation of the structure imposed by the stop sign is negotiated during the activity. Things may run smoothly at the intersection - but there will also be interruptions. An impatient driver piggybacks on the driver in front of him. A pedestrian decides to ignore the stop sign altogether.

The first part of this paper will develop the notion of a coordinating representation in the context of the interdisciplinary literature on coordination. The second part focuses on the cognitive engineering task of building coordinating representations for computer-mediated joint activities. The last part of the paper presents an experimental evaluation of the utility and function of the coordinating representation.

The Problem of Coordination

Whether it is greeting someone, or planning a potluck dinner party, or moving through a doorway, or forming a queue at the coffee shop - there are always problems of coordination. When you greet someone, depending on the circumstance, you may say “hi", shake hands, slap hands, hug, kiss, or ignore. Each form of greeting (except the last) requires coordination (and cooperation) among the participants. For a potluck dinner party, the meal must be coordinated for taste, balance, and variety. The meal can include appetizers, main courses, desserts, and beverages; a preponderance of one or the other detracts from the meal. For many doorways, there is not enough room for two people (say, in conversation) to pass through the doorway shoulder-to-shoulder. To effectively move through the doorway the participants must coordinate on an order as to which one passes through the doorway first, second, ... and who is to hold the doorway open. The queue at the coffee shop begins and ends at a certain place; people line up in the order they arrive.

Some examples of coordination problems are the assignment of roles, the establishment of location, manner, and structure, and issues of sequencing; timing and co-reference.

Suppose Tipper and Al are re-arranging furniture in the house. Each of the above kinds of coordination problem may come into play as they move the old couch from the living room, down the stairs, around the corner, through a doorway into the basement. Al's role may be to back down the stairs holding the front of the couch; Tipper walks forward holding the backend of the couch. Initially they meet in the living room. Their path as they carry the couch begins in the living room and ends at the basement. Their manner may be slow and cautious, so as to avoid bumping into walls and doorways. At certain points they are tilting the couch at an angle so they can move down the stairwell without bumping the couch into the ceiling. Coordination at *the boundaries between phases of the activity* (Clark, 1996), must be jointly engineered by Tipper and Al as they shift from moving down the stairs to moving through the doorway. In order to move the couch down the stairs, Tipper and Al need to establish co-references for features of the stairwell (e.g., the low ceiling) or the situation (e.g., an unexpected problems they encounter). Some of the coordination problems are 'solved' before

action begins (e.g., Al walks backwards and Tipper walks forwards); others are resolved as the action proceeds (e.g., the coordination problems entailed by the low ceiling in the stairwell).

The term *structure for behavior* is used here to refer to the kinds of information exchanged between Tipper and Al in order to achieve their joint task and maintain coordination - examples of which are the assignment of roles, the path, the manner… Not all the information exchanged is a structure for the current behavior. For example, Tipper and Al are also socializing as they proceed with their activity. Nor are all structures for joint behavior exchanged at runtime: both Tipper and Al are likely to have prior experience at moving a couch through a doorway. Using both the social exchanges of information about structure and the recollection of prior related experiences, the participants must jointly reason out and construct a behavior which achieves their shared goal of moving the old couch from the living room to the basement.

The structures relevant to a given act in the current activity that are available before the act may be either recalled, planned, the result of an explanation, or designed. Both Tipper and Al may remember previous occasions when they moved furniture. For the difficult portions of their task, they may explicitly create a shared plan (Grosz & Sidner, 1990), an agreed to structure - you do this and I'll do that - for the behavior. If the structure for behavior is produced after a given behavior is completed, it is called an explanation (Mitchell, et. al., 1986), which can become realized in future related episodes of joint activity. Over time, for joint activities that Tipper and Al regularly do, behaviors become conventionalized and designs for the structure of those behaviors will begin to emerge (Alterman & Garland, 1998).

As Tipper and Al perform their activity, the fact that they are co-present allows them to monitor the progress of their joint activity. Because they can see one another, they can use body position to communicate information. Throughout their activity they can speak to one another in order to co-develop, for example, a procedure for moving the couch down the stairway. Their comments to one another are exchanged without delay, in the course of their joint behavior. The actions that form their conversation and activity occur sequentially.[1]

Other kinds of joint activity do not allow for a face-to-face interaction, so other methods or mediums must be introduced to support the exchange of structural information. Performance depends on the participants communicating - by these altered means - information relevant to design, plan, and commitment. For computer-mediated tasks, the trick will be to convert structures (designs) that are naturally produced in conversation by the users into external representations that can mediate similar sorts of cooperative activities in the future. The design of the external representations that are developed will focus on simplifying the most difficult coordination problems that typically confront users.

The Coordinating Representation

A coordinating representation is an external representation shared among participants in a joint activity. It is designed for the activity-at-hand and reduces the complexity of the coordination task. It mediates and structures the activity. It has the designated purpose of helping participants to achieve coordination in non face-to-face cooperative activities. Its meaning is based on conventional interpretation. It signals to the participants - without dictating action - that a convention of behavior is in place.

Consider the scene at the airport. For the passenger, the printed itinerary that her travel agent sent her helps her to stay coordinated. The itinerary identifies her flight destination and number. When she arrives at the airport, she uses the listed flight number to select among the flights and gates listed on the departure monitor for American Airlines. The design of the destination monitor (first listed in alphabetical order of destinations and then by time of departure) reduces her cognitive load in finding the departure gate for her flight. When it comes to finding her departure gate, the itinerary and the departure monitor are two coordinating representations that help to replace a face-to-face interaction with a mediated one.

Alternately, suppose the passenger needs to "check in" some luggage before proceeding to her gate. What coordinating representations are used to insure her bag makes the trip? Now, upon arrival at the airport, the passenger looks for the check-in counter for the airline from which she purchased her ticket. Large signs displaying airline logos indicate where each airline is located. Smaller signs divide the queue into first class and regular passengers. As the passenger puts her bag on the scale, the clerk attaches a tag indicating airline, flight destination, and flight number. Later, a bagger must transport in a truck the bags to the cargo space of the plane. A *complex sheet* that links flights to destinations and unique aircraft identification numbers is used by the bagger to achieve his goal (Goodwin & Goodwin, 1996). The organization of the complex sheet makes the access of information more efficient.

Each of the coordinating representations used to get both the passenger and her luggage on the correct plane has both a social and an individual function. From the perspective of the social, the coordinating

[1] This list is adapted from an analysis developed by Clark & Brennan (1991) to explicate differences among various kinds of mediated communication.

representation preserves a set of references for objects shared among the participants. From the perspective of the individual, the coordinating representations simplify access to the information that is being exchanged.

There are many other examples of coordinating representations in everyday life. An appointment slip helps a patient to return to the dentist's office on the right day at the right time. A mail order catalogue helps the customer and the sales office reach agreement on purchase items, sizes, and prices. Tax forms help to coordinate citizens and IRS personnel in their efforts to exchange information....

Experimental Platform: VesselWorld

For the last several years we have been building a same time/different place groupware system (VesselWorld) as an experimental platform for analyzing real time computer-mediated collaborations. A demo of the system was run at CSCW 2000 (Landsman et. al., 2000).

There are several important characteristics of the joint activity of participants in a VesselWorld problem-solving session. Participants have different roles (both predefined and emergent). Cooperation and collaboration are needed to succeed. Participants must develop a shared understanding of an unfolding situation to improve their performance. Uncertainty at the outset makes pre-planning inefficient in many circumstances. There are numerous problems of coordination.

In VesselWorld, there are three users engaged in a set of cooperative tasks that require the coordination of behavior in a simulated environment. In the simulated world, each participant is a captain of a ship, and their joint task is to find and remove barrels of toxic waste from a harbor. Two of the users operate cranes that can be used to lift toxic waste from the floor of the harbor. The third user is captain of a tugboat. The cranes are able to individually lift and carry small or medium toxic waste barrels, jointly lift large barrels, and jointly lift (but not carry) extra large barrels. The tugboat cannot lift barrels, but can attach to, and move, small barges. Small barges may hold multiple barrels. Each captain has a small radius of perception. Many barrels require the use of other equipment in addition to the cranes. The tugboat captain is the only one who can examine barrels to determine equipment needs. Barrels can be leaking - or will begin to leak if they are dropped - in which case the leak must be contained by the tug.

The VesselWorld interface provides to each user several different windows of information. The World View (not shown) depicts the harbor from the point of view of a participant, who can only see a limited region at one time. The World View graphically represents several kinds of information about the location and status of objects from the perspective of an individual participant. A second window of information is used for planning. A third window allows a user to access more detailed information about visible objects. A chat window allows participants to communicate with one another using an electronic chat.

In a base version of the VesselWorld system, participants can only coordinate by electronic chatting. Most of the participant dialogue is centered on the barrels, and how effort can be coordinated in removing the barrels from the harbor and transporting them to a large barge. During a problem solving session, the flow of information between participants using the base system is continuous. It is the responsibility of each actor to add information conveyed to him by another actor to his or her private representation (either by taking notes, marking the map, or remembering), or be prepared to examine the history of chatting at some appropriate future time. Any information that is lost, misunderstood, never recorded, or never transmitted in the first place, can lead to discrepancies between the participants' individual assessments of the situation.

An analysis of participant dialogue determines a set of problem areas in organizing behavior in relation to a shared domain object. So, for example, a large volume of information must be exchanged over the naming, status, location, and properties of the toxic wastes. In a second version of the system, coordinating representations are introduced that basically structure and simplify the exchange of information in the problem areas of coordination.

Analysis of Electronic Chatting

The electronic chatting amongst participants is used as a basis for developing some coordinating representations. As the analyst reviews the discourse, she needs to look closely at using coordinating representations to simplify the most common interactions, fix repeated errors in coordination, and replace conventions developed by users during the course of a problem-solving session. The goal is not to entirely replace other forms of communication with coordinating representations. Rather the analyst wants to use coordinating representations to improve performance - thereby simplifying the interaction - at critical points in the ongoing cooperation among participants.

The analysis was framed by cognitive work on the problem of coordination that was presented at the beginning of the paper. Figure 1 shows a list of the kinds of methods that were used by participants to coordinate their joint activities. The participants did some planning by assigning roles or agreeing to sets of actions. During the activity, a fair amount of chatting was used to initiate joint actions that were tightly coupled; for example, to lift an extra large waste, the cranes have to begin lifting during the same time

segment. Also found in the discourse were examples of the participants creating conventions to simply the exchange of information for recurrent problems of coordination. Chatting was continuously used throughout each session to establish references and exchange information about shared domain objects.

- Plan (provide orientation: delimit tasks)
  - Plan to do; Role assignments
- During Activity (Entry & Exit into Phases)
  - Synchronization; sequencing; step; turn-taking; Action taken; See; Initiating Statement
- Develop conventions
- Co-Referencing and the exchange of information
  - Refer to status, location, feature, identity of object

Figure 1: Taxonomy of coordination methods.

Figure 2 shows a sample dialogue of the kind of close coordination users needed to do in order to time closely coupled activities. At 1 and 2, after jointly lifting a large barrel, Crane1 and Crane2 agree to do a joint carry followed by a joint load onto a barge. It will take three moves to reach their destination. In lines 3, 4, and 5, they tell each other they submitted their first move. At 8 the tug suggests a convention to simplify coordination. At 9 and 10, Crane1 and Crane2 tell each other they are ready to do the second part of the move. At 14, Crane1 states she is doing the third move. At 15-18 they plan, and then they submit actions, to do the joint load. At 21 and 22, they celebrate.

1. Crane1: now a joint carry, clicked at 375,140 got 3 carrys
2. Crane2: i will do same
3. Crane2: move to first location
4. Crane1: submitted first
5. Crane2: ditto
6. Crane1: again?
7. Crane2: yes
8. Tug1: do you want to just type something in after submitting each turn
9. Crane1: submitted second
10. Crane2: ditto
11. Tug1: just some shorthand or something, for everyone so we know whats going on
12. Crane1: submitted third
13. Tug1: submitted
14. Crane2: submitted third
15. Crane2: Crane1: load, and then i'll to the same
16. Crane1: submitted load
17. Crane2: ditto
18. Tug1: submitted move
19. Crane2: hey, i think that worked!
20. Crane1: looks like it's Miller time. I think we did it.

Figure 2: A sample dialogue

Three Coordinating Representations

The analysis of the pilot study discourse identified three recurrent areas of coordination activity:

1. Timing of closely coupled activities
2. Establish references for, and exchange information about, shared domain objects and their status.
3. Higher-level planning to manage multiple cooperative activities

None of these should be surprising as possible areas of difficulty: each of these has been suggested by prior theoretical analysis. But there are also other potential problem areas. So the problem for the cognitive engineer is to determine which things are problematic for the task-at-hand.

Some sketches of three coordinating representations were developed and later refined through an interview with one of the test groups in the pilot study.[2] After the interview, the iterative design process continued by a cycle of (re)design, implementation, and evaluation. The periodic evaluation came in several forms, including expert reviews, in-group experimentation, and study groups paid for at Brandeis University. What resulted from this process were three coordinating representations that were designed both to simplify the interaction among participants (the social part) and structure it so as to reduce the cognitive load of each user (the individual part) in her use of the mediating representation.

The coordinating representation showed in Figure 3 allows a user to compare his projected actions to those of the other participants. The next few projected steps of each actor is displayed in a labeled column for each participants. The actions are listed in order from top to bottom. (So, the next projected step of Crane1 is to do deploy equipment and then he will lift some waste.) Each user has control of only one column, his/her own. This representation improves timing on exit and entry of phases for tightly coordinated phases of activity by allowing participants to compare each other's next few projected actions.

Figure 3: Timing of joint actions.

The second coordination representation is the *object list,* which contains a list of objects with relevant

[2] S. Kirschenbaum at NUWC collected the data for this pilot study.

properties in a table format. Columns provided information about the name, object type, location, and equipment needed for a given object. The organization of this information reduces the cognitive load for the individual, by organizing information relevant for decision making into predetermined representational structure.

A third coordinating representation was designed to allow the users to do *high-level planning*. The idea was to create a space where the participants could rapidly sketch a high-level plan that would help them to manage multiple open tasks. There are three columns in this window: one for each actor. Each column could be used, for example, to abstractly represent that each actor is currently searching a different part of the harbor. Further down each column, the participants could indicate that they are committed to a plan to move, in order, wastes 1, 2, and 3 onto a small barge. A palette at the top the window allows users to rapidly build a description of a joint action sequence. Actions are one of a small set of action primitives, i.e., MOVE, SEARCH, and CONTAIN. Color-coding of entries in the high-level plans allows participants to indicate both accomplished tasks and future commitments.

Experimental Evaluation

An experimental evaluation conducted at Brandeis compared the performance of teams of participants with (and without) the coordinating representation. Three groups could only electronically chat during problem-solving sessions, and three groups could chat but also had access to coordinating representations. Each team was trained and then played for about 10 hours over several sessions of problem solving. All events that occur during a problem-solving session are recorded in a log file by the system. A VCR-like device was used to review and analyze the decision making of each group. A more complete discussion and detailed analysis of the experimental data, with numerous examples, can be found in Alterman et al (2001).

Quantitative Analysis

One measure of general performance is the amount of clock time it took the participants to solve a problems:[3] there was a 49% improvement in clock time to complete task for those groups using coordinating representations. Another measure of user work indicates that there was a 38% reduction in the number of events generated while completing tasks of comparable difficulty. Because the coordinating representations pre-structure certain exchanges of information we expected to see a reduction in the quantity of electronic chats: there was a 57% reduction in the amount of electronic chatting. Because one of the coordinating representations dealt with commitment (high-level planning), another with timing, and a third with the exchange of information about equipment requirements for lifting barrels, we expected to see a reduction in domain errors: total errors were down 61%. However, a closer analysis of the data reveals that the high-level planning coordinating representation was used hardly at all. Further discussion of this last point is below.

[3] These results have 95% confidence intervals and are normalized for the complexity of the problem. Problem complexity is a weighted sum over all wastes taking into account size, equipment needed, and distance from large barge for each waste.

Qualitative Analysis

For the groups that did not have access to coordinating representations, the predominant method for maintaining a common view of the world was for the participants to continuously *report* on their current activity via electronic chatting. One strategy for avoiding differences in assessment was to engage in a conversation to *review* the status of one or another of the shared domain objects. Whenever discrepancies in the assessment of a situation unexpectedly developed the participants engaged in *repair* work to re-mediated between alternate representations of "reality". Participants also regularly *confirmed* that somebody else's report or repair was received. Each of these techniques was important to the functioning of the groups using the basic system in maintaining a joint sense of their common enterprise. These groups also developed additional structures to simplify the exchange of information using the electronic chat window. The simplest of these were naming conventions. A second example was a set of conversational structures that were developed by each group to support coordination of closely coupled actions.

The general advantage of the coordinating representation was that it simplified the problem of establishing a consistent representation of the situation among the participants.

One advantage that accrued to the users who had access to the coordinating representation that supports the *timing of joint actions* is that it required no extra work on the part of the participants to build. In order to submit an action to the system the users needed to add it to their "plan" anyway. So, from the point of the view of the users who have access to the shared planning window, having to talk about their cooperative activity is just extra work. Another advantage was that one actor now has the opportunity to spot potential problems in another actor's plan.

Much of the dialogue that accompanied the discovery of a new waste in the groups using the basic system was mediated by the *object list* for the groups that had

access to coordinating representations. Identifiers were attached to each of the "objects" that were found. Pointing and clicking was used to add entries to the object list, thus precise locations for each of wastes that were found could be stored. These aspects of the object list simplified the process by which the actors established references and referents. Because the object list was a shared representation, much of the consistency checking that the users of the base system had to engage in was no longer necessary. Rather than having two private representations that must periodically be reconciled by electronic chatting, the users could share a single representation. This scheme reduced the number of conflicts between different conceptions of the shared workspace, but it also eliminated the work involved in re-mediating discrepancies between alternate views of the shared domain objects.

The high-level planning window was not used by any of the groups. The surveys we collected from the subjects show that the chief problem with the high-level planning windows was that, given the rewards it provided, it required too much work to complete. Further analysis shows that the problems that the high-level planning window was designed to fix continued to occur.

We are developing two solution paths to fixing this problem. The first is to do a better job of modeling the work of the individual user in cooperation with the other users (Feinman & Alterman, 2001). The second approach is develop some AI techniques that would allow the system to fill out portions of the high-level planning window semi-automatically (Introne & Alterman, 2000).

Concluding Remarks

The overarching interest of this research is to continue to develop a framework for Cognitive Science that depends not only on the mental operations of the individual but also on the social interaction within which it is embedded. (An underlying thesis is the cognition is irreducibly social.) Application domains involving the computer-mediation of joint activity are significant areas of research because they allow one to investigate both the social and individual aspects of cognition. The methodology that was used for developing coordinating representations for VesselWorld reflects these commitments and attitudes. There are two parts to the methodology: a social and an individual one. During the social part, the developer can collect data on the usage of the base system and do an analysis of the information exchanged among participants (a discourse analysis) that helps them to stay coordinated. A key is to identify recurrent problems of coordination that showed up in the pilot version of the system. During the individual part, the designer tunes the initial approximations for coordinating representations to the cognitive operations of the individual user. During this phase, representations are iteratively designed to simplify the work of the individual user in creating and accessing the coordination information that is shared among the participants.

Acknowledgments

This work was supported in part by ONR Contracts N00014-96-1-0440 and N66001-00-1-8965.

References

Alterman, R. and Garland, A. (1998). Convention in Joint Activity. TR# CS-98-199, CS Department, Brandeis University (1998). To appear in *Cognitive Science* 25(4), 2001.

Alterman, R., Feinman, A., Landsman, S. and Introne, J. Coordination of Talk: Coordination of Action. TR# CS-01-217.

Clark, H. (1996). *Using Language*. Cambridge University Press.

Clark. H. and Brennan, S. (1991). Grounding in Communication. In Resnick, Levine, and Teasly editors, *Perspectives on Socially Shared Cognition* (pp. 127-0149)

Feinman, A., and Alterman, R. (2001) Modeling Communicative Behavior in a Groupware System. TR CS-00-211 Computer Science Department, Brandeis University. To appear in HCI 2001.

Goodwin and Goodwin (1996). Seeing as situated activity: Formulating planes. In Engeström and Middleton editors, *Cognition an communication in work.* Cambridge University Press, Cambridge, U.K.

Grosz, B. and Kraus, S. (1996). Collaborative plans for complex group action. *Artificial Intelligence*, 86:269-357.

Grosz, B. and Sidner, C. (1990). Plans for discourse. In Cohen, P. R., Morgan, J., and Pollack, M. E., editors, I*ntentions in Communication*, pages 417-444. Bradford Books, Cambridge, MA.

Introne, J. and Alterman, R. (2001) Segmenting Usage Data in Collaborative Systems. TR CS-01-215. Appeared in workshop on "Dealing with Community Data" at CSCW 2000.

Landsman, S., Alterman, R., Feinman, A, Introne, J., VesselWorld and ADAPTIVE, TR CS-01-213, Brandeis University. Demo given at CSCW 2000.

Lewis, D. (1984) *Convention: A philosophical study.* Harvard University Press.

Mitchell, T. Keller, R. and Kedar-Cabelli, S. (1986) Explanation-based generalization: A unifying view. *Machine Learning,* 1:47-80.

Sacks, H., Schegloff, E., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking in conversation. *Language*, 50:696-735.

An Integrative Approach to Stroop: Combining a Language Model and a Unified Cognitive Theory

Erik M. Altmann (ema@msu.edu)
Department of Psychology
Michigan State University

Douglas J. Davidson (doug@eyelab.msu.edu)
Department of Psychology
University of Illinois at Urbana-Champaign

Abstract

The rich empirical puzzle of the Stroop effect has traditionally been approached with narrowly focused and somewhat atheoretical models. A recent exception is a simulation model based on the WEAVER++ language theory. The present model, WACT, combines components of WEAVER++ with the memory and control processes of the ACT-R cognitive theory. WACT accounts for the time course of inhibition from incongruent word distractors, facilitation from congruent word distractors, the lack of effect of color distractors, and the semantic gradient in inhibition. WACT goes beyond WEAVER++ to account for Stroop performance errors as well as latencies, and its implementation in a unified cognitive theory opens doors to broader coverage of Stroop phenomena than standalone models are likely to attain. Documented and executable code for WACT is available for inspection and comment at **www.msu.edu/~ema/stroop**.

Introduction

The Stroop effect is the mental confusion (and its behavioral consequences) induced when a word such as green is printed in a color such as red and the task is to name the color (red, in this case). Word meaning (green, in this case) seems to be processed automatically, in some sense, causing it to interfere with the color-naming task. Thus, the system may think green even though it sees red, because it can't stop itself from reading the word.

The rich pool of data on the Stroop effect (see MacLeod, 1991) has to date been approached with relatively lean cognitive theory. For example, the dominant simulation models remain the connectionist models of Cohen, McClelland, and Dunbar (1990) and Phaf, Van der Heijden, and Hudson (1990). The former model shows that Stroop phenomena can be simulated with simple information-processing units appropriately wired together. However, it makes no obvious contact with other cognitive theory – there are no identifiable linguistic or perceptual constraints, for example. Also, the model fails to capture the time course of inhibition, in which inhibition falls off gradually as the distractor occurs further ahead of the target (Glaser & Glaser, 1982; Glaser & Glaser, 1989; Sugg & McDonald, 1994). Indeed, simulated interference increases monotonically with temporal separation (Cohen et al., 1990, Fig. 7), suggesting basic flaws in the model's representation. The SLAM model (Phaf et al., 1990) is embedded in a theory of visual attention, but says little about the role of memory and executive control, and fails to capture the time course of inhibition (their Fig. 14a) and the asymmetry of reading and naming (their Fig. 14b).

Our approach to modeling Stroop effects is to integrate existing theory from other cognitive domains. Our model adopts mechanisms of the WEAVER++ language theory (Roelofs, 2000c), which explains Stroop phenomena in terms of competing lemmas (syntactic properties of words; Roelofs, 2000a, 2000b). These linguistic mechanisms are integrated into the ACT-R cognitive theory (Anderson & Lebiere, 1998), which specifies memory and executive-control mechanisms. The resulting model, which we refer to as WACT, goes beyond WEAVER++ to account for errors as well as latencies, and benefits from its embedding in ACT-R in terms of potential extensions to other phenomena. ACT-R suggests how automaticity of the dominant Stroop task might develop (MacLeod & Dunbar, 1988), and implements a theory of perceptual, motor, and cognitive constraints (Byrne & Anderson, in press) that could integrate a diverse range of Stroop effects into one model.

We begin by describing the effect to be explained – the time course of Stroop inhibition, in which latency and errors increase as distractor onset approaches target onset (Glaser & Glaser, 1989). We then describe WACT and its account of these effects, as well as its account of Stroop facilitation, a semantic gradient, and the non-effect of color distractors. In the discussion, we examine WACT's limitations and some possible extensions suggested by ACT-R.

The Time Course of Stroop Inhibition

Figure 1 illustrates the Stroop effect of primary interest here. The empirical data (solid lines) are from Experiment 1 of Glaser and Glaser (1989), in which a word and a color are shown with some temporal separation. Of interest here is the case in which the word (appearing first) is the distractor and the color patch (appearing second) is the target (the stimulus to which the participant responds). Thus, the word green might precede the color red by 100 msec. This temporal difference is the stimulus onset asynchrony (SOA). By convention, SOA is negative when the distractor precedes the target.

The latency difference measure in Figure 1 is derived by subtracting neutral latencies from incongruent and congruent latencies. On neutral trials, the distractor is a stimulus that consists of letters but is not a color word (e.g., xxxx). On incongruent trials, the distractor is a color word whose meaning conflicts with the color patch (e.g, green and red). On congruent

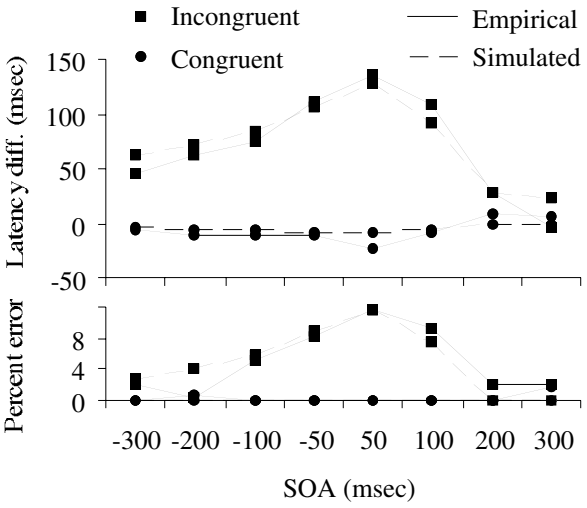


Figure 1: Stroop inhibition and facilitation. Latency difference is Incongruent/Congruent minus a neutral condition (see text). Stimulus onset asynchrony (SOA) is target onset minus distractor onset. Empirical data are from Glaser and Glaser (1989), Exp. 1, and simulated data are from WACT.

trials, the distractor is a color word whose meaning matches the color patch (e.g., red and red). In all three kinds of trials, the target stimulus is the color patch.

In Figure 1, the upper curves in each panel (square markers) are from incongruent trials, in which the distractor interferes with the target. As is typical, interference is greatest (i.e., the latency difference is greatest) when SOA is near zero – when target and distractor appear at roughly the same time. Note that even if the distractor occurs slightly after the target (e.g., an SOA of 50 msec), it still causes substantial interference. The lower curves in each panel (round markers) are from congruent trials, in which the "distractor" actually slightly facilitates performance.

The error measure in Figure 1 is the raw percentage of substitution errors, or trials on which the wrong response word was given. (With no detectable facilitation from congruence, there is no need for a difference measure.) Only incongruent naming is particularly error prone, and there, as with latency, interference is greatest at near-zero SOA.

The WACT Model

Long-term lexical knowledge in WACT is organized in a multi-layer declarative network, as shown in Figure 2. The top level of this network contains semantic nodes, or concepts. Below this is the lemma layer, which contains syntactic information (lemmas) crucial for fitting a word into the grammatical organization of a phrase or sentence. Below the lemma layer is the form layer, which contains the information necessary to produce an individual word.

In WEAVER++ and WACT, interference and facilitation occur at the lemma layer. Word stimuli have direct access to their lemmas, whereas non-verbal

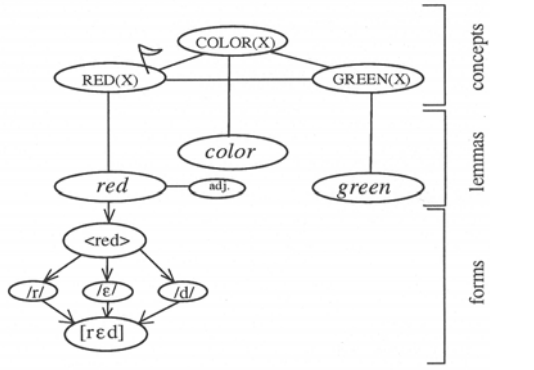


Figure 2: WEAVER++ long-term lexical knowledge (from Roelofs, 2000a). WACT represents concepts, lemmas, and the top layer of forms (e.g., <red>).

stimuli like colors gain access only indirectly, via concepts (Figure 3). The direct link from a word stimulus to its lemma is the route by which words trigger automatic language processing. The benefit of this automaticity is efficiency, helping the system to meet immediacy constraints on comprehension (Just & Carpenter, 1987). The cost of this automaticity, on our view, is that it leaves behind traces of information that can interfere with subsequent tasks like color naming.

A word stimulus automatically activates the corresponding lemma. We assume that the purpose of this activation, relative to language comprehension, is to facilitate parsing of subsequent tokens. For example, the stimulus "and" might establish an expectation for a subsequent conjunct, by virtue of causing the "and" lemma to be active when the conjunct arrives. Stroop interference (and facilitation) are caused by the activation of a lemma for the distractor word. If this distractor lemma is incongruent with the target lemma, it produces a form of response competition when the system tries to retrieve the target lemma. On the other hand, if the distractor lemma is congruent with the target lemma, then the system benefits from intrusions of the distractor lemma.

In WEAVER++ and WACT, latency to retrieve a target depends on the target's activation relative to distractors – the more active the target is relative to its distractors, the quicker it is retrieved. Relative activation is a common way to formalize interference (Baddeley & Hitch, 1993; Luce, 1959; Murdock, 1985; Neath, 1993). In ACT-R the formulation is

$$P_i = \frac{e^{A_i/s}}{\sum_j e^{A_j/s}} \qquad \text{Equation (1)}$$

where P_i is the probability of retrieving item i on a given attempt given the j items in memory at the time. A_i is the activation of i, and s is system noise.

Importantly, WACT (unlike WEAVER++) specifies the processing consequences of retrieving the wrong item on a given attempt. WEAVER++ predicts latency simply by scaling relative activation. WACT actually uses the retrieved lemma to decide how to respond, so

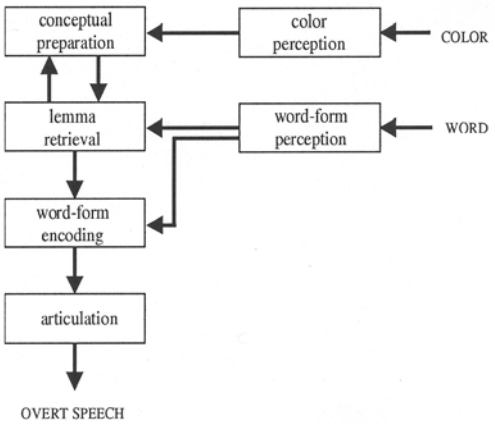


Figure 3: Stages of linguistic processing in WEAVER (from Roelofs, 2000a) and in WACT.

there is the possibility of an incorrect lemma retrieval causing a substitution error.

The lemma-retrieval process is shown in Figure 4. The figure is taken from Murdock (1974), but retrieve-decide models like this are common (e.g., Anderson & Bower, 1972; Kintsch, 1970; Watkins & Gardiner, 1979) and map naturally to ACT-R memory-retrieval productions. Such a process also explains tip-of-the-tongue effects, in which subjects appear to monitor correctness of retrievals (e.g., Levelt, 1983). Probed with a word stimulus, the system tries to retrieve the corresponding lemma. Any retrieved lemma is evaluated (in the decision process) by comparing the current concept to conceptual cues retrieved with the lemma. In case of a mismatch, the system tries again. Eventually the system retrieves a lemma it considers correct, or runs out of time. Either way, the last lemma retrieved is the basis for form retrieval (the next stage of language production; Figure 3). If the last lemma retrieved is incorrect, then form processing begins with the wrong input, likely causing a performance error. Thus, the distractor lemma interferes with the target lemma by affecting the duration and potentially the output of the lemma-retrieval process.

In WACT, the amount of interference caused by a distractor lemma depends on its activation, which in turn depends on the time elapsed since the distractor stimulus was presented. Activation in ACT-R is

$$A = \ln(\frac{2n}{\sqrt{T}}) \qquad \text{Equation (2)}$$

where n is the number of times the item has been retrieved and T is the length of the item's lifetime. For a distractor lemma, T=n=1 when the distractor stimulus is presented. After that n remains essentially constant, but T increases throughout the trial, causing A to decrease (decay). Thus, the more time elapses between distractor and target, the more the distractor lemma decays and the less it intrudes on the target lemma.

To illustrate, Figure 5 shows activation values from Monte Carlo simulations of naming trials at various SOAs. The top curve is the activation of long-term lemma representations. For these representations, n and

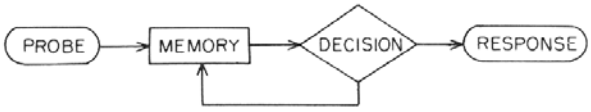


Figure 4: The WACT retrieve-decide process for lexical retrieval (from Murdock, 1974).

T are both large, so activation is stable over short intervals. The bottom curve is activation of the distractor lemma. At large negative SOAs, the distractor lemma decays by the time the target appears, but at near-zero SOAs its activation is close to that of long-term lemmas. Syntactic representations are known to decay rapidly (e.g., Potter & Lombardi, 1990; Sachs, 1967), and here this rapid decay explains the time course of Stroop interference. On an incongruent trial, the system must retrieve a long-term lemma in order to process the target correctly, and this retrieval is faster and more accurate if the distractor lemma has decayed.

Two other comments on Figure 5 are in order. First, the activation values reflect the sum of two sources: base-level activation (A in Equation 2) and associative activation from cues like the current stimulus; these are the two possible sources of activation in ACT-R (Anderson & Lebiere, 1998). The sum of these two sources is the activation factor in the item's likelihood of being retrieved (A in Equation 1). That said, the scale on the ordinate of Figure 5 is arbitrary, because relative activation, not absolute activation, is what governs retrieval probability in WACT. A second point is that the lower curve ends at SOA 100. This means only that at SOA 200 and 300, the target lemma was always retrieved before the distractor appeared. When the target is retrieved in time to avoid interference from the distractor, the model implementation simply skips the step of activating the distractor lemma, as the trial is functionally over by then.

Comparing WACT to Data

WACT behavioral data, from the same simulations that produced the activations in Figure 5, are presented in Figure 1 (dashed lines). The fits are quite respectable: r^2=.98 and RMSD=11.0 for latencies, r^2=.94 and RMSD=1.4 for errors. The model clearly captures the peak in inhibition near zero SOA and the gradual falling-off (leftward) as the distractor word is presented

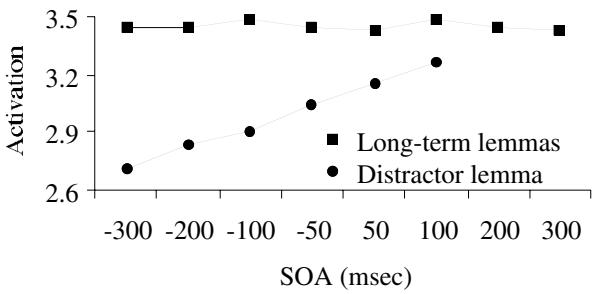


Figure 5: The time course of activation in WACT, showing decay of distractor lemma at negative SOAs.

further ahead of the target color. Inhibition also falls off sharply for positive SOAs (rightward), where the target is usually fully processed and the response formulated by the time the distractor appears.

WACT also captures Stroop facilitation, a common though relatively small effect. In the model, facilitation arises when a congruent distractor lemma is correct, in which case the "distractor" lemma is indistinguishable from the target. Functionally, activating a congruent distractor lemma is equivalent to a slight increase in activation of the target lemma.

An important "non-effect" captured by WACT is that a color distractor has no effect on word reading, either inhibitory or facilitative. In the model, color distractors have no effect because color stimuli are not processed ballistically, as words are. In response to a color stimulus, the system does not automatically activate a lemma; in terms of cognitive economy, there is no reason to process an arbitrary stimulus verbally unless the task requires it. Thus, a color distractor leaves behind no activated lemmas to interfere with the subsequent target word. This account is a point of distinction between WACT and WEAVER++, which accounts for the non-effect of colors not in terms of ballistic processing but in terms of levels of processing (Roelofs, 2000b; personal communication).

The Semantic Gradient

Beyond time course effects, another important effect captured by WACT is the semantic gradient, in which a distractor like "lawn" inhibits the naming of a target color like red. This inhibition arises presumably because lawn primes some representation of green, which then conflicts with the response to red. This effect is important because it is one of a class of effects involving higher-level semantic relations among categories (Glaser & Dungelhoff, 1984; Roelofs, 2000c). It is also an opportunity to compare WACT and WEAVER++, and to address a point of contention in how best to model gradient effects in ACT-R.

In WACT, the semantic gradient arises from operation of the retrieve-decide process (Figure 4) at the concept level. Earlier, we described this process operating at the lemma level, but in fact the process operates at each of the levels of speech production (Figure 3), as befits a general process for detecting and correcting memory errors. Thus, just as with lemma retrieval, when the model needs a concept, interference from incorrect concepts will degrade performance.

To illustrate how the semantic gradient arises, suppose the distractor word is lawn and the target color is red. When WACT sees "lawn", the lawn concept is activated as a side effect of processing the lawn lemma. The corresponding assumption in WEAVER++ is that activation spreads from lemma to concept in parallel as it spreads from lemma to form (Figure 3). In WACT, the lawn concept cues related knowledge through semantic priming. Among the concepts related to lawn is green – which also belongs to the response set for the current task (i.e., sometimes the target color is green). The combination of priming from the lawn concept and priming from the task environment is enough to cause the green concept to intrude occasionally on the concept for the actual target color. That is, having processed lawn, the model may think green, even if it sees red. Relevant data (from Exp. 5 of Glaser & Glaser, 1989) appear in Figure 6. The semantic gradient is represented by the small, positive latency difference across SOAs, reflecting modest interference from distractors like lawn. (The small peak at SOA 50, which Glaser & Glaser, 1989, attribute to random variation, is unrelated to the semantic gradient.) WACT again follows the trend, with distractors like lawn causing some interference but not as much as distractors like green.

The WACT account of semantic gradients may be another point of distinction relative to WEAVER++. In ACT-R, activation spreads only one link from whatever cues are in the focus of attention. Thus, spreading activation over a distance of multiple links requires a sequential process of chained retrievals in which each retrieval brings the next cue into the focus of attention. In WEAVER++, by contrast, activation spreads uncontrolled throughout the lexical network. Though attenuated by distance (number of intervening links) from the activation source, this uncontrolled spreading seems to make WEAVER++ quite sensitive to representational assumptions. For example, current reports (Roelofs, 2000a, 2000b) suggest that activation from the lawn word would reach the green lemma (via the concepts lawn and green), causing conflict with the red lemma. The same reports suggest that activation from the lawn word would also reach the red lemma (via the concepts lawn, green, and red), compensating for the activation reaching the green lemma. Thus, the word lawn could produce inhibition, facilitation, or neither with naming the color red, depending on the relative strengths of the various associations involved.

The WACT account of semantic gradients is also important because it shows that such effects can be accommodated by ACT-R's core theoretical premises. ACT-R assumes that performance (including memory performance) adapts to the statistical structure of the environment (Anderson & Lebiere, 1998; Anderson &

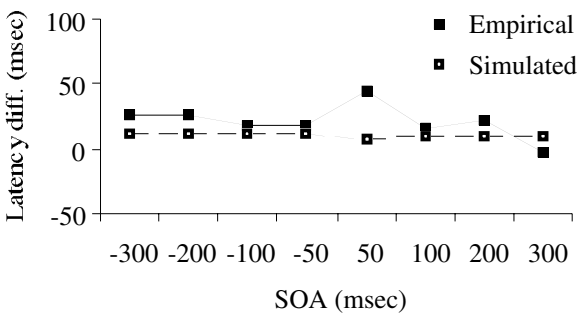


Figure 6: Inhibition from distractors like "lawn". Empirical data are from Glaser and Glaser (1989), Exp. 5, Cond. 2. Simulated data are from WACT.

Milson, 1989). Thus, WACT assumes that lawn and green concepts are associated in memory because they co-occur in the environment. Associative mechanisms also account for temporal gradients in order memory (Altmann, 2000). Nonetheless, the adequacy of such representations has been questioned, and ACT-R has come to incorporate a "partial matching" mechanism for fitting gradient data (e.g., Anderson & Matessa, 1997). The current work suggests that this mechanism, which has no clear motivation in terms of independent theoretical constraints, is best viewed as a simplifying assumption and not as a part of ACT-R theory proper.

Model Parameters

The parameters used to fit the data in Figures 1 and 6 were as follows. Activation noise (set at 0.33), or *s* in Equation 1, causes some retrieval attempts to produce the incorrect target. Encoding noise (0.0205) causes some stimuli to be encoded out of order at small SOAs. (Activation and encoding noise both index logistic variance; see Anderson & Lebiere, 1998.) The limit on retrieval attempts (3) affects how soon the retrieve-decide process gives up and outputs its last retrieval. Other parameters affect associative activation spreading from a cue. High strength (8.9 units of activation) applies to perceptual cues and medium strength (6.9 units) applies to mental cues. Low strength (4.9 units) applies to lawn as a cue for green, so is relevant only to the fit in Figure 6.

Discussion

Stroop effects are complex and diverse and it seems clear that broad coverage will elude us as long as we continue to approach them with standalone models. A comprehensive theory is required, in which the interactions of various cognitive subsystems can be simulated to investigate whether particular Stroop phenomena emerge as a natural consequence. Several directions indicated by the marriage of WEAVER++ and ACT-R in WACT are discussed below.

First, there is the question of Stroop development – how a process like reading becomes automatic enough to interfere with other tasks like color naming. MacLeod and Dunbar (1988, Experiment 3) demonstrated that this development can be induced through training. Their participants received extensive practice (daily, for a month) on associations between arbitrary shapes and color names. These associations eventually became automatic enough to interfere with color naming, which, before training, had been the more automatic task. Would ACT-R, as a unified theory that integrates learning mechanisms, allow WACT to be extended to develop automaticity?

A WACT account of automaticity would likely be grounded in ACT-R's utility-based theory of procedural skill acquisition. Procedural skills in ACT-R are represented as production rules that govern retrieval from declarative memory. Skill acquisition itself is represented in part as the acquisition of cost-benefit knowledge about individual production rules – the more a rule succeeds, the more it is preferred when the system has a choice. This mechanism has been used to account for set effects (Lovett, 1998) on the view that these are driven by frequency of rule use. In WACT, automaticity is represented by fixed settings of these cost-benefit parameters. That is, productions that read a word stimulus have high utility and thus are preferred to productions for less-used skills like color naming. It seems feasible and useful to extend WACT to simulate training data sets like that of MacLeod and Dunbar (1988). With such an extension in hand, one could assess its predictive value by manipulating system parameters like the rate at which skill acquisition takes place.

A second question concerns the relationship between Stroop phenomena and task switching. Stroop interference and executive control interact, in that switching to controlled tasks like color naming is easy, whereas switching to automatic tasks like reading is hard (Allport, Styles, & Hsieh, 1994). However, Stroop conflict is robust to task uncertainty – inhibition effects are largely unchanged when the task is determined dynamically on each trial by stimulus order (Glaser & Glaser, 1989). These disparate effects of task may inherit explanations from studies of task switching conducted within ACT-R (Altmann & Gray, in press; Sohn, Ursu, Anderson, Stenger, & Carter, 2000).

A third question concerns the relationship between Stroop phenomena and the psychological refractory period paradigm used to investigate perceptual, motor, and cognitive bottlenecks. The data in Figure 7, from which the latency-difference measure in Figure 1 was computed, hint at a bottleneck in Stroop processing. The slowing near SOA zero suggests a "jamming" of some kind (Meyer & Kieras, 1997) when stimuli appear close together in time. WEAVER++ accounts for this effect (Roelofs, personal communication) in terms of its activation dynamics. In contrast, ACT-R incorporates a structural and processing theory of bottleneck effects generally (Byrne & Anderson, in press). Extending WACT in this direction would take it well beyond current Stroop models, by integrating perception, action, language, memory, and executive control in one running model.

Finally, we hope to extend WACT to phrase

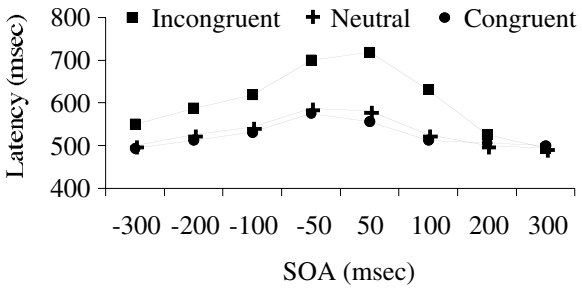


Figure 7: Latencies from which the difference scores in Figure 1 were computed, showing that processing slows near zero SOA, regardless of condition.

production, in particular the production of conjunctive phrases. When the task is to name the color and read the word, utterance onset depends on which response is to be given first. In particular, utterance onset is delayed when the color is to be named first (and the word meaning is incongruent). These data help to characterize planning in speech production, but also offer an opportunity to integrate Stroop phenomena more broadly with psycholinguistic theory.

Acknowledgements

Thanks to Tom Carr, Ardi Roelofs, and the conference reviewers for comments on the original draft.

References

Allport, A., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In Umilta & Moscovitch (Eds.), *Attention and performance IV* (421-452). Cambridge: MIT Press.

Altmann, E. M. (2000). Memory in chains: A dual-code associative model of positional uncertainty. *Proc. 3rd international conference on cognitive modeling* (9-16). Veenendaal, NL: Universal Press.

Altmann, E. M., & Gray, W. D. (in press). Forgetting to remember: The functional relationship of decay and interference. *Psychological Science*.

Anderson, J. R. & Bower, G. (1972). Recognition and retrieval processes in free recall. *Psych. Rev., 79*, 97-123.

Anderson, J. R., & Lebiere, C. (Eds.). (1998). *The atomic components of thought*. Hillsdale: Erlbaum.

Anderson, J. R., & Matessa, M. (1997). A production system theory of serial memory. *Psych. Rev., 104*, 728-748.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psych. Review, 96*, 703-719.

Baddeley, A. D., & Hitch, G. (1993). The recency effect: Implicit learning with explicit retrieval? *Memory & Cognition, 21*, 146-155.

Byrne, M. D., & Anderson, J. R. (in press). Serial modules in parallel: The psychology refractory period and perfect time-sharing. *Psychological Review*.

Cohen, J. D., McClelland, J. L., & Dunbar, K. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review, 97*, 332-361.

Glaser, M. O., & Glaser, W. R. (1982). Time course analysis of the Stroop phenomenon. *Journal of Experimental Psychology: HPP, 8*, 875-894.

Glaser, W. R., & Dungelhoff, F.-J. (1984). The time course of picture-word interference. *Journal of Experimental Psychology: HPP, 10*, 640-654.

Glaser, W. R., & Glaser, M. O. (1989). Context effects in Stroop-like word and picture processing. *Journal of Experimental Psychology: General, 118*(1), 13-42.

Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension*. Boston: Allyn and Bacon.

Kintsch, W. (1970). Models for free recall and recognition. In D. A. Norman (Ed.), *Models of Human Memory*. New York: Academic Press.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, *14*, 41-104.

Lovett, M. (1998). Choice. In Anderson & Lebiere (Eds.), *Atomic components of thought* (255-296). Hillsdale, NJ: Erlbaum.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109*(2), 163-203.

MacLeod, C. M., & Dunbar, K. (1988). Training and Stroop-like interference: Evidence for a continuum of automaticity. *J. Exp. Psych.: LMC, 14*, 126-135.

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance. *Psych. Rev., 104*, 3-65.

Murdock, B. B. (1974). *Human memory: Theory and data*. New York: Wiley.

Murdock, B. B. (1985). An analysis of the strength-latency relationship. *Mem. & Cog., 13*, 511-521.

Neath, I. (1993). Distinctiveness and serial position effects in recognition. *Mem. & Cog., 21*, 689-698.

Phaf, R. H., Van der Heijden, A. H. C., & Hudson, P. T. W. (1990). SLAM: A connectionist model for attention in visual selection tasks. *Cognitive Psychology, 22*, 273-341.

Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language, 29*, 633-654.

Roelofs, A. (2000a). Attention to action: Securing task-relevant control in spoken word production, *Proc. 22nd annual meeting of the Cognitive Science Society* (411-416). Mahwah, NJ: Erlbaum.

Roelofs, A. (2000b). Control of language: A computational account of the Stroop asymmetry. *Proc. 3rd int'l conference on cognitive modeling* (234-241). Veenendaal, NL: Universal Press.

Roelofs, A. (2000c). WEAVER++ and other computational models of lemma retrieval and word-form encoding. In L. Wheeldon (Ed.), *Aspects of language production*. Philadelphia: Psychology Press.

Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception and Psychophysics, 2*, 437-442.

Sohn, M.-H., Ursu, S., Anderson, J. R., Stenger, V. A., & Carter, C. S. (2000). The role of prefrontal cortex and posterior parietal cortex in task switching. *Proc. National Academy of Sciences, 97*(24), 13448-13453.

Sugg, M. J., & McDonald, J. E. (1994). Time course of inhibition in color-response and word-response versions of the Stroop task. *Journal of Experimental Psychology: HPP, 20*, 647-675.

Watkins, M. & Gardiner, J. (1979). An appreciation of generate-recognize theory of recall. *Journal of Verbal Learning and Verbal Behavior, 18*, 687-704.

Age of Acquisition in Connectionist Networks

Karen L. Anderson (kanders@cs.ucsd.edu)
Garrison W. Cottrell (gary@cs.ucsd.edu)
Computer Science and Engineering Department 0114
Institute for Neural Computation
University of California, San Diego
La Jolla, CA 92093-0114 USA

Abstract

Recently, there has been a resurgence of interest in the role of the Age of Acquisition (AoA) of an item in determining subjects' reaction time in naming words, objects, and faces. Using the number of epochs required to learn an item as a direct measure of AoA in connectionist networks, Smith, Cottrell & Anderson (in press) have shown that AoA is a stronger predictor of final Sum Squared Error than frequency. In this paper, we replicate Smith *et al*. using more realistic frequency distributions for the items, and examine why some patterns may be learned earlier than others. First, we have found that the same patterns tend to be learned early and late by networks differing in their initial random weights; hence, the issue is, what property of the patterns determines AoA? We have found that even very weak pattern similarity structure is a strong predictor of AoA when frequency is controlled for. Also, we have found evidence that such a similarity structure may still be an important factor in determining AoA even when pattern frequency is varied.

Introduction

Ever since Carroll & White (1973) reanalyzed Oldfield & Wingfield' s (1965) naming latency data and discovered that frequency was not significant when AoA was considered, controversy has surrounded discussions of the import of the two variables. Technological and methodological refinements have led to agreement that both frequency and AoA play significant roles. Hence, interest has returned to the pursuit of understanding the mechanisms underlying AoA effects.

It had been proposed recently (Morrison & Ellis, 1995; Moore & Valentine, 1998) that connectionist networks would be incapable of exhibiting AoA effects because training on late patterns would cause "catastrophic interference" resulting in the unlearning of early patterns. However, this sort of interference is only found if training on early patterns ceases. Ellis & Lambon Ralph (2000) demonstrated AoA effects in a neural network by training the net on an "early" set of patterns and then simply adding a second set of "late" patterns halfway through training.

Smith et al. (in press) independently demonstrated AoA effects in networks. In contrast to the staging method of Ellis & Lambon Ralph (2000), where AoA is assumed to correspond to the time at which patterns are presented to the network (early or late), all patterns were presented to the model from the outset. AoA can then be *measured* for each pattern individually as the time during training when the pattern is learned. Using this more natural definition, Smith et al. reported significant effects of AoA on naming latency (defined as the residual error on a pattern after training is completed, a measure of the network's *uncertainty*).

What we would like to know is why certain patterns are learned earlier than others, and how early learning of a pattern comes to affect the network' s performance. Ellis & Lambon Ralph's (2000) approach cannot be used to find out why patterns are acquired in a particular order as it *imposes* an order by staging pattern presentation. Instead, we vary properties of the patterns and then measure AoA, as in Smith et al. Ellis & Lambon Ralph (2000), do suggest why early AoA is important for final performance – the network is more "plastic" earlier in training, so items that are learned first have the opportunity to make the biggest impression on the weights.

We also want to know whether our finding that AoA is a stronger predictor of final error than frequency survives a more realistic version of Zipf's (1935) frequency distribution than was used by Smith et al. Here we show that it does.

Methods

Our investigation is organized around a series of experiments in which we replicate and extend network simulations and analyses previously reported by Smith et al. (in press). We begin with one of the simplest connectionist models of lexical access, an autoencoder network. This kind of network simply reproduces its input on its output through a set of hidden units, and has seen surprisingly wide application in cognitive modeling. We then extend our simulations to more complex mappings.

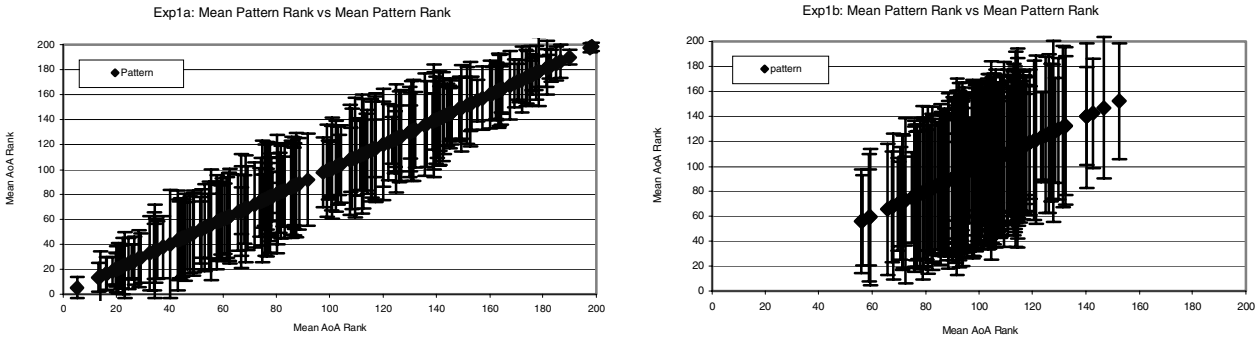


Figure 1: Comparison of pattern AoA variance between experiments using the same pattern set for all simulations (left) and 10 different pattern sets (right).

Table 1: Average correlations between pattern similarity measures and AoA.

	AoA	Mean Cosine	Mean R^2	Density	Mean Distance
AoA	1.0000				
Mean Cosine	0.0487	1.0000			
Mean R^2	-0.4751	-0.0399	1.0000		
Density	0.0806	0.9886	-0.0412	1.0000	
Mean Distance	0.1069	-0.3058	-0.0437	-0.1825	1.0000

Experiment 1

Smith et al. (in press) report finding a strong correlation between AoA and SSE in their first experiment in which they trained ten autoencoders on the same set of equally frequent patterns. Training all networks on the same set of patterns ignores the possibility that the order in which the patterns in the set will be acquired by the network may depend on some property of the training set. To examine this possibility, we replicated Smith et al.'s first experiment in two ways: first, using the same set of randomly generated patterns (an exact replication) and, second, using a different pattern set for each network. The first replication allowed us to perform an analysis of the AoA rank order correlation of patterns between pairs of networks – if networks trained on the same pattern set tend to acquire patterns in the same order then the rank order correlations between pairs of networks should be significant, implicating a property of the training set in driving acquisition order. The second replication allowed us to see whether Smith et al.'s finding concerning the correlation between AoA and SSE maintained across a larger set of patterns.

Methods For the first replication, ten groups of ten networks were trained with all of the networks in a group using the same pattern set. In the second replication, a single group of ten networks were trained with each network using a different pattern set. For both replications, the pattern sets consisted of 200 randomly generated 20-bit patterns in which each bit had a 50% chance of being on. All networks were 20-15-20 autoencoders trained via backpropagation, using learning rates of .001, momentums of .9, and initial random weights between 0.1 and -0.1. All patterns were presented every epoch. Training was continued until 98% of the patterns were acquired (where "acquired" means its SSE went below 2.0). The AoA of a pattern was taken to be the first epoch in which it was acquired.

Results Smith et al. (in press) reported a correlation coefficient of 0.749 between SSE and AoA averaged over all 10 networks. For both replications, we found similar mean correlations: 0.773 (0.038) and 0.756 (0.050), for a randomly chosen group in the first (same pattern) replication and the group in the second (different patterns) replication, respectively. Thus, our replication supports the finding of Smith et al. that AoA and SSE are strongly correlated.

Although we arrived at that same result in the first replication, our second replication does not support the conclusion that AoA is independent of properties of the training set. Our examination of the AoA rank order correlation between groups of networks trained on the same pattern set revealed that networks trained from different initial weights tend to learn the patterns in a set in a similar order. The pair-wise AoA rank order correlations between networks in the same group averaged over all pairs in all groups (N=450) were found to be 0.485 (σ=0.061), using Kendall's τ, and 0.665 (σ=0.071), using Spearman's ρ. Figure 1 illustrates this relationship. Both graphs

in the figure plot the mean AoA values for each pattern on both axes. The graph on the left is for 10 networks in one of the groups in the first replication using the same pattern set for each network. In order to estimate how chance behavior would look, we simply aligned the different pattern sets used in the second simulation based on pattern numbers and, in the graph on the right in Figure 1, we plot the means and standard deviation for all patterns with the same number. Note how the means in the plot on the left do not cluster about the center as do those in the plot on the right, and that those on the left have smaller standard deviations.

Having found that *some* property of the training set contributes to the AoA of the patterns in the set, our next goal was to attempt to identify what that property might be. Note that in choosing random 20 bit patterns, we are selecting vectors randomly from a 20 dimensional space. Since the maximum number of vectors that can be mutually orthogonal in such a space is 20, and we are selecting 200 vectors, each vector in the set will necessarily be closer to some vectors in the set than to others. This unavoidable clustering of patterns in the vector space is what we refer to when we speak of the similarity structure of a randomly chosen set of training patterns. Since the patterns are randomly chosen, the average pair-wise correlation between patterns is small (0.0581, σ = 0.0056 for an exemplary set), but non-zero.

For one of the pattern sets used in the first replication, we computed for each pattern the mean cosine, R^2, and Euclidean distance between the pattern and all others, and the pattern density (% bits "on"). The correlations between these measures and the patterns' AoA values were computed for each network and then averaged together. As Table 1 shows, the negative mean R^2 between a pattern and all others in the network is on average the best predictor of the pattern's AoA. We performed a repeated measures multiple regression analysis (Lorch & Myers, 1990) using mean R^2, mean Euclidean distance and density as predictors of AoA, and found that the null hypotheses that the mean regression coefficients are equal to 0 can be rejected with $p<<0.000001$, $p=0.000013$, and $p=.020976$ for mean R^2, mean Euclidean distance, and density, respectively. Thus, we are led to believe that the small and subtle structure reflected by the inter-pattern correlations among the patterns in even a randomly chosen set has a strong role in determining the order in which those patterns will be learned.

Experiment 2

In this experiment we again replicate and extend Smith et al. (in press). Like Smith et al., we aim to show that AoA effects persist in our model when frequency is added as a variable, and to compare the strengths of these effects to those found in human studies. We improve upon Smith et al. by first, using more realistic frequency distributions and second, by examining the role the shape of the frequency distribution has on the relative contributions of frequency and AoA to naming latency.

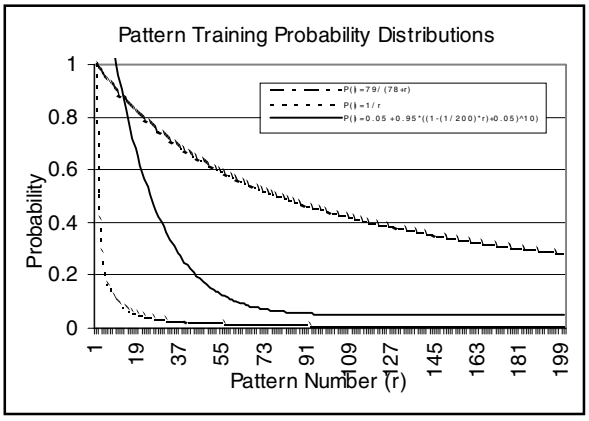


Figure 2: Comparison of training probability distributions.

Methods We again use ten autoencoders with differing pattern sets, but we vary the frequency of presentation of the patterns within each set. In manipulating pattern frequency, we aim to simulate the well-known Zipf distribution, where a small number of words occur very frequently – that is, the frequency of a word is proportional to the reciprocal of the word's frequency rank. We took two approaches to simulating this distribution. In the first approach, we randomly assign ranks to patterns and train on each pattern with probability 1/rank in each epoch. In the second approach, we take account of the fact that the most frequent words tend to be function words (like "a", "the", "and", etc.) and that human naming studies seldom use such words. Hence, a more accurate model of the frequency distribution of words used for naming stimuli should start lower on the Zipf curve. In order to determine a reasonable starting point, we needed to make an estimate of the frequency ranking of the most frequent word likely to be used in a naming study. To do so, we examined the Celex database (Baayen, Piepenbrock & Gulikers, 1985), and found the rank of the most frequent noun with an imageability rating of 500 or greater in the MRC Psycholinguistic database (Coltheart, 1981). The 500+ imageability criterion was chosen somewhat arbitrarily (the mean rating for words in the MRC database is 450), but was intended to find roughly where concrete nouns show up on the Zipf's curve. The noun selected by this process was "man" with a rank of 78. Hence, our second replication of this experiment randomly

Table 2: Network and human naming study correlation data.

	Networks			Object Naming				Word Naming
	Smith et al.	1/r	79/(78+r)	E&M	S&Y	BM&E	C&W	M&E
r(aoa, sse)	0.749	0.727	0.763	0.626	0.683	0.700	0.77	?
r(logf,sse)	-0.730	-0.462	-0.324	-0.405	-0.456	-0.455	?	-0.388
r(aoa,logf)	-0.283	-0.259	-0.212	-0.377	?	?	?	?
r(log-aoa, sse)	?	0.755	0.826	?	?	?	?	0.244
r(log-aoa, logf)	?	-0.524	-0.273	?	?	?	?	-0.414

(E&M = Ellis & Morrison, 2000; S&Y = Snodgrass & Yuditsky, 1998; BM&E = Barry, Morrison & Ellis, 1997; C&W = Carroll & White, 1973; M&E = Morrison & Ellis, 2000)

assigned a rank, r, between 1 and 200 to each pattern and then presented that pattern with probability (79/78+i) for training on each epoch.[1]

Our central motivation for using more than one frequency distribution in this (and subsequent) experiments is to determine how the shape of the distribution might influence the relative contributions of AoA and frequency to SSE. We hypothesized that training with a frequency distribution from the beginning of Zipf's curve would tend to make the frequency of a word a stronger determinant of its final SSE than would training with a distribution that started after the curve began to flatten. We were also interested in verifying the results obtained by Smith et al., since they used only a single pattern set and just a Zipf-*like* frequency distribution. In particular, they presented pattern r for training each epoch with a probability given by:

$P(r) = 0.05 + 0.95*((1-(1/200)*r)+0.05)^{10})$

Figure 2 shows a graph comparing all three distributions. Note that the Smith et al. distribution has many more "high-frequency" words than does the 1/r distribution, and that it spans a larger range of probabilities than does the 79/(78+r) distribution.

Results Table 2 shows the correlation coefficients obtained from the three network models, as well as regression coefficients obtained from human object and word naming studies. The results show that the network model correlations look much more similar to the object naming data than to the word naming data. This is a bit counter-intuitive given that the networks are being trained to autoencode — word naming is a less arbitrary mapping than object naming and, hence, seems like a better match to the autoencoding task. The results of our next experiment suggest a reason for this discrepancy. We put off further discussion until then.

The main difference between the network models' correlations is that Smith et al.'s r(logf,sse) is much greater than both the other two network models and the human data. Examining Figure 3, we might suppose that the 79/(78+r) distribution has a weaker frequency effect than Smith's distribution since the frequency differences among patterns are not as pronounced. The 1/r distribution may be weaker than Smith's for a similar reason — while it covers a maximal range of frequencies like Smith's, it has relatively few at the high frequencies and, so, little differentiation in terms of frequency for the vast majority of its patterns. As both the models with 1/r and 79/(78+r) are closer to the human data than the model with Smith's curve, though both are at somewhat opposite extremes in terms of frequency distribution, support is lent to the notion of using a true Zipf based distribution. Furthermore, the slight weakening of the effect of frequency on SSE in the 79/(78+r) model compared to the 1/r model suggests that the choice of data set used in human naming experiments (object names will not be at the top of the Zipf curve) could influence the observed strength of correlation between naming latency and frequency and, possibly, explain some of the differences in findings reported in these studies.

Experiment 3

Having demonstrated AoA effects in the presence of frequency in networks trained to perform an autoencoding task, Smith et al. (in press) then examined how different levels of consistency in the mapping task represented by the pattern set influenced AoA and frequency effects. While spelling to sound is reasonably consistent mapping, spelling to meaning or faces to names are not. Again, we were interested in replicating Smith et al. to see whether their results still held when using the more realistic 1/r and 79/(78+r) frequency distributions and unique training pattern sets for each simulation.

Methods The networks were modified compared to the previous experiments in order to make learning the less consistent pattern sets possible: the number of hidden units was increased to 50, and the objective function was changed from SSE to cross-entropy.

[1] The scale factor of 79 is used only to minimize the number of epochs required for learning the set – it simply guarantees that the most frequent pattern is presented exactly once every epoch, while the relative frequencies of the patterns remain unchanged.

Ten pattern sets were randomly created, as before. From this set of ten, eleven sets of ten were created by randomly flipping bits of the target patterns with eleven levels of probability evenly distributed between 0.0 and 0.50 − in 100% consistent pattern sets, the target patterns were exactly the same as the input patterns (autoencoding); in the 0% consistent case, each bit in the target pattern had a 50% chance of being flipped from the input setting (a completely random mapping, like sound to meaning).

Results The graphs in Figure 3 plot for each level of mapping consistency the mean correlation coefficients and the mean coefficient p values of multiple regressions (N=10 for each point) on network SSE with AoA and frequency as the independent variables. The plots reveal that AoA is a stronger and more significant predictor of naming latency than is frequency in our model across all levels of consistency. As we previously noted, these charts may help explain why the data from Experiment 2 look more like object naming than word naming. Even though word naming is a more consistent mapping than object naming, it is still not 100% consistent, as was the task used in Experiment 2. From the graphs of variable significance on the bottom in Figure 3, it is obvious that the case of 100% consistency is somewhat of a discontinuity, resembling 0% consistency more than it does 90% consistency. Autoencoding is not a good model of word naming tasks.

Experiment 4

We view the mean pair-wise AoA rank order correlation between simulations trained using the same pattern set as a measure of the contribution of pattern set similarity structure to determining the order in which words are acquired. In analysis of the AoA effects observed in the networks of experiment 1, we computed this measure for several groups of simulations and found it to be significant. Since experiment 1 was concerned only with autoencoding networks, we wondered whether the effect of pattern structure has as much influence on pattern AoA in networks trained to perform less consistent mapping tasks. We were also curious as to whether the order in which patterns were presented for training would have much effect on the ordering of AoA among the patterns. To answer these questions, we designed our final experiment.

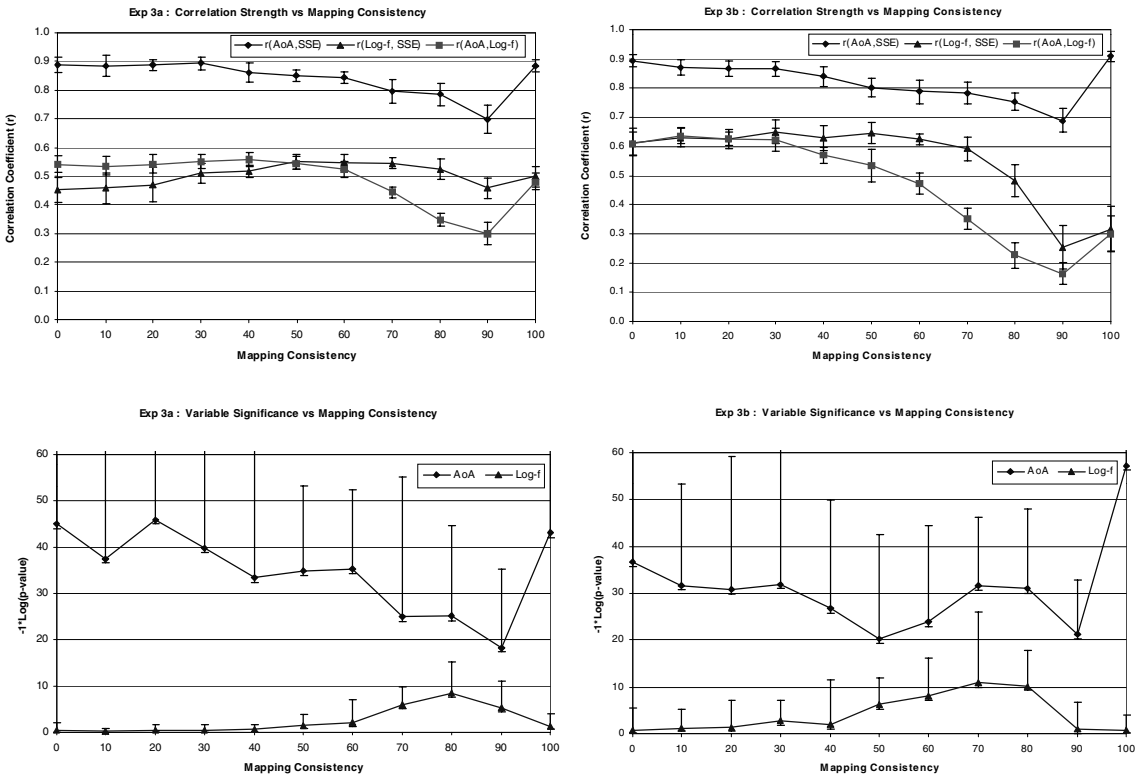


Figure 3: Comparison of the effect of consistency on correlation strength (top) and significance (bottom) between models trained with a 1/r frequency distribution (left) and a 79/(78+r) frequency distribution (right).

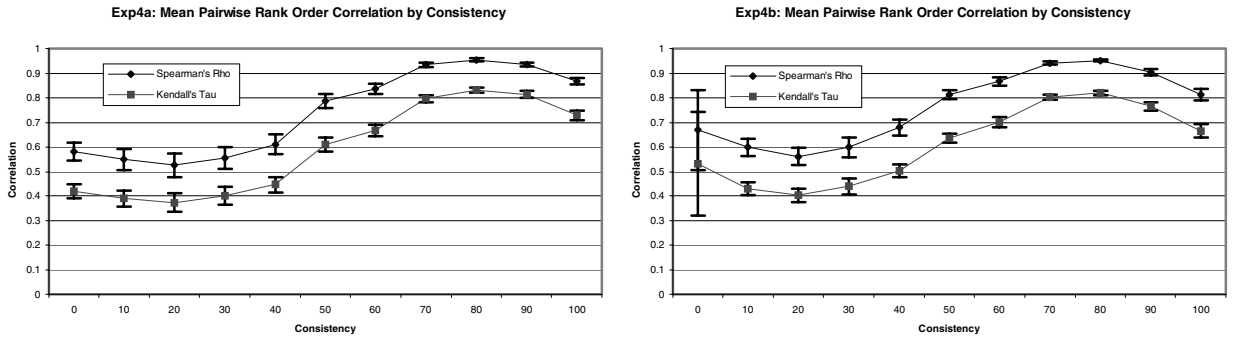


Figure 4: Comparison of pairwise pattern AoA rank order correlations across consistency levels between sets of networks trained with (right) and without (left) randomized pattern presentation order.

Methods One pattern set was arbitrarily selected from each consistency group used in experiment 3, to create a set of eleven training sets with varying levels of consistency ranging between 0% and 100%. For each level of consistency, two sets of ten networks were trained from different random initial weights. The first set was trained with every pattern presented for training in the same order on every epoch. For the second set of ten, all patterns were presented in a new and random order each epoch. Because we were interested only in observing the influence of pattern set similarity structure across training set consistency levels, all patterns in all sets were trained with a uniform frequency distribution. Otherwise, the networks were the same as those in experiment 3.

Results The graphs in Figure 4 plot the mean rank order correlations for each level of consistency. They reveal that, not only is pattern set similarity structure important at all levels of consistency, but that it is also mostly independent of pattern presentation order. The one notable difference between random and non-random presentation ordering occurs at 0 consistency, showing up as a large standard deviation in the plot on the right in Figure 4. This experiment also reveals that similarity structure is generally more influential on AoA at higher levels of consistency.

Conclusion

We have shown that the similarity structure among items is an important determinant of AoA across a variety of mapping tasks. Future work will concentrate on more realistic similarity structures within the domains and ranges of the mappings, such as similarities between words, between faces, and between meanings. On the issue of frequency vs. AoA, the regressions performed in experiment 3 reveal that AoA is a stronger predictor of naming latency in our models than frequency. While AoA and frequency are clearly correlated, there appears to be a fundamental effect of an item becoming encoded in the network weights before other items. Frequency may be the key, but AoA is the door to performance.

Acknowledgments

We wish to thank GURU, Rich Golden, Mark Smith and Dave Noelle for their contributions and support.

References

Barry, C., Morrison, C. M., & Ellis, A. W. (1997). Naming the Snodgrass and Vanderwart pictures: Effects of age of acquisition, frequency and name agreement. *QJEP*, 50A, 560-585.

Baayen, R. H., Piepenbrock R. & Gulikers, L. (1995). *The CELEX Lexical Database (CD-ROM).* Phila., PA: U. of Penn., Linguistic Data Consortium.

Carroll, J. B. & White, M. N. (1973). Word frequency and age of acquisition as determiners of picture-naming latency. *QJEP*, 25, 85-95.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *QJEP*, 33A, 497-505.

Ellis, A. W. & Lambon Ralph, M. A. (2000). Age of Acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *JEP:LMC*, 26(5), 1103-1123.

Lorch, R.F., Jr., & Myers, J.L. (1990). Regression analyses of repeated measures data in cognitive research. *JEP:LMC*, 16, 149-157.

Morrison C. M. & Ellis, A. W. (2000). Real age of acquisition effects in word naming and lexical decision. *British J. of Psychology*, 91(2), 167-180.

Oldfield, R. C. & Wingfield, A. (1965). Response latencies in naming objects. *QJEP*, 17, 273-281.

Smith, M. A, Cottrell, G. W., and K. L. Anderson (in press). The early word catches the weights. To appear in *NIPS* 12. Cambridge, MA: MIT Press.

Snodgrass, J. G., & Yuditsky, T. (1996). Naming times for the Snodgrass and Vanderwart pictures. *Beh. Res. Meth., Instr. & Comp.*, 28, 516-536.

Zipf, G. K. (1935). *The Psycho-Biology of Language.* Boston, MA: Houghton Mifflin.

The Processing & Recognition of Symbol Sequences

Mark W. Andrews (mwa1@cornell.edu)
Department of Psychology; Uris Hall
Ithaca, NY 14853 USA

Abstract

It is proposed that learning a language (or more generally, a sequence of symbols) is formally equivalent to reconstructing the state-space of a non-linear dynamical system. Given this, a set of results from the study of nonlinear dynamical systems may be especially relevant for an understanding of the mechanisms underlying language processing. These results demonstrate that a dynamical system can be reconstructed on the basis of the data that it emits. They imply that with minimal assumptions the structure of an arbitrary language can be inferred entirely from a corpus of data. *State-Space reconstruction* can implemented in a straightforward manner in a model neural system. Simulations of a recurrent neural network, trained on a large corpus of natural language, are described. Results imply that the network sucessfully recognizes temporal patterns in this corpus.

Introduction

Complex pattern recognition is often characterized by means of a simple geometric analogy. Any object or pattern may be described as a single point in a high-dimensional space. For example, a square grayscale image that is 256 pixels in length, may be described as a point in the 256^2 dimensional space of all possible images. A collection of such images is a set of points in this space. If these patterns are not entirely random, this set will reside in a subspace of lower dimensionality. To learn the structure of these images, an organism or machine must discover a compact parametric representation of this subspace. This might take the form of, for example, finding a reasonably small set of basis vectors that will span the subspace and projecting each image onto these vectors. Having done this, each image can be classified in terms of a new and more meaningful coordinate system. You effectively describe 'what is there' in terms of 'what is known'.

This geometric approach is routinely employed in the study of visual object recognition, but may easily be extended to a wide range of categorization and classification tasks. In almost all cases, however, the patterns under study have been multi-dimensional *static* patterns. In contrast, the study of *temporal* pattern recognition using this or related approaches has not been well-developed. For example, one of the most widely employed techniques for temporal pattern recognition, Hidden Markov Models are limited in their generality due to their fundamental inability to handle patterns above a certain complexity. This absence of general models for temporal pattern recognition is evident in the study of human language processing, which traditionally has eschewed serious consideration of statistical learning and pattern recognition.

This paper aims to introduce a general framework for the study of temporal pattern recognition. This is developed in the context to language processing, but it could be extended in a straightforward manner to most other cases of temporal patterns. First, a general characterization of the problem of language learning and language processingis proposed. Then, some recent results in the study of nonlinear dynamical systems are described. These are seen as being especially relevant for an understanding of the mechanisms underlying temporal pattern recognition, especially with regard to language processing. Finally, simulations with a recurrent neural network are described, which suggest successful pattern recognition of English sentences.

The processing of symbol sequences

A paradigm for the study temporal pattern processing, especially language processing, has developed as a result of the deep relationship between formal languages and abstract automata (Chomsky 1963)[1]. Any language (or more generally, any sequence of symbols), can be described as the product of a particular automaton. By this account, learning a language is equivalent to identifying a particular automaton on the basis of a sample of the language that it generates. More formally, an automaton A is specified by the quadruple $\langle X, Y, F, G \rangle$. X and Y are sets known as the state and the output spaces, respectively. The functions $F : X \mapsto X$ and $G : X \mapsto Y$

[1]The correspondence between formal languages and abstract automata can be summarized by the so-called Chomsky hierarchy: Classes of automata that are increasingly restrictive versions of the Turing machine produce classes of languages described by increasingly restrictive generative grammars. The *regular* languages R are produced by strictly finite automata, the *context-free* languages CF are produced by pushdown stack automata, the *context-sensitive* languages CS are produced by linear bounded automata and the *recursively enumerable* languages RN are produced by unrestricted Turing machines. $R \subset CF \subset CS \subset RN$, and likewise for their corresponding automata.

are the state-transition and the output functions, respectively. Beginning at time t_0 and continuing until t_∞, the sentence-generator A constantly changes from one state in X to the next, according to its state-transition function F. At each transition, a symbol from the set Y is emitted, according to its output function G.

A langage learner attempts to identify the nature of this automaton on the basis of a sample of the language that it generates. That is to say, the language learner is exposed to a finite sequence of Y and from this must attempt to identify $A = \langle X, Y, F, G \rangle$. Having attained knowledge of A, the learner is said to have full knowledge of the structure of the language. The learner has the capability to produce all the sentences of the language, including the infinite number of sentences that were never seen. Likewise, the learner has the capability to parse the syntactic form of any sentence of the language. This ability also extends to the infinite number of never-encountered sentences. As syntactic parsing is a necessary precondition for the interpretation of language, it is said that the language has been learned once knowledge of its grammar has been attained.

While the correspondence between formal languages and automata has allowed the problem of language learning to be given an explicit characterization, these automata $A = \langle X, Y, F, G \rangle$ have always been taken to be discrete parameter systems. To continue this paradigm it is useful to demonstrate the correspondence between generative grammars and continuous as well as discrete automata. Within nonlinear dynamical systems theory, the study of *symbolic dynamics* has made apparent the relationships between formal languages, generative grammars and continuous dynamical systems. Symbolic dynamics refers to the practice of coarse-coding the ambient state-space of a dynamical system into a finite set of subspaces and assigning a symbol to each. Whenever the system enters a partition, the assigned symbol is emitted. In this way, the trajectories of the dynamical system can be represented as strings of symbols. Unless the system is entirely stochastic, only a certain subset of strings will occur. It can be shown that these strings define a language and the system producing them can be described by a generative grammar (Bai-Lin & Wei-Mou 1998).

The relationship between languages, grammars and dynamical systems has been further described by Tabor (1998). In that work, and in Tabor (2000), the computational capacities of a pushdown stack automaton were identified with those of a stochastic dynamical system, based on an *iterated function system*. This was used to demonstrate the recognition of context-free languages by a simple $2-$dimensional dynamical system. Following Tabor's approach, it is reasonable to propose that any language (or any symbol sequence) generating process may be legitimately described as a *continuous* as well as a *discrete* system. Accordingly, and by keeping a strict analogy with the automaton $A = \langle X, Y, F, G \rangle$ described above, it is possible to introduce the corresponding continuous system, A' defined by the quadruple $\langle \vec{x}, y, f, g \rangle$.

By introducing $A' = \langle \vec{x}, y, f, g \rangle$, the language generating process is being explicitly defined as a nonlinear dynamical system. For example, the system may be described by a set of coupled differential equations

$$\dot{x}_i = f(\vec{x}, \delta)$$

where $\vec{x}$ is the system's state and $\dot{x}_i$ is a vector-field defined on an m-dimensional manifold M. δ is a unspecified stochastic element in the system. The *language* being produced by this system is a result of the coarse-coding function

$$y = g(\vec{x}),$$

where y is a variable representing the *symbols* of the language. However, there are still formal similarities between the discrete automaton $A = \langle X, Y, F, G \rangle$ and its continuous counterpart $A' = \langle \vec{x}, y, f, g \rangle$. $\vec{x}$ is the state-space of A' and y is a variable representing its output. The function $f\colon \vec{x} \mapsto \vec{x}$ describes the state evolution of the system while $g\colon \vec{x} \mapsto y$ is an output function. In fact, the only essential distinction between $A = \langle X, Y, F, G \rangle$ and $\mathcal{A} = \langle \vec{x}, y, f, g \rangle$ is that in the latter case the state-space $\vec{x}$ is continuous, rather than discrete, and the evolution of the system described by f is smooth, preventing discontinuous leaps through space.

State-Space Reconstruction

A language learner can be said to be attempting to identify the process generating the language. If this generating process is described as a continuous dynamical system, the objective is to model the dynamical system $\dot{x} = f(\vec{x})$ on the basis of the language it outputs, $y_{t_0}, \ldots, y_t$. Prima facie, this problem is widely intractable. The symbols to which the learner is exposed do not identify the state of the system. They are a product of the composition of two unknown and probably non-linear functions, f and g. However, it may be fruitful to consider the analogy between this problem and a more general problem encountered in the experimental analysis of complex systems. For example, a scientist observing a sequence of individual measurements from a complex physical process (e.g. a fluid in turbulent motion) may be interested in understanding the properties of the underlying system. In the absence of prior knowledge and without loss in generality, the system can be taken to be a stochastic dynamical system, whose functional form is completely unknown. The scientist must infer its functional form on the basis of the measurement data alone. One of the more remarkable outcomes of dynamical systems theory is that in many general cases this problem is tractable. In virtue of the analogy, the manner by which this is done may also elucidate the problem of language learning.

Packard, Crutchfield, Farmer & Shaw (1980) were first to demonstrate that a dynamical system could be reconstructed entirely on the basis of its output. They proposed that *any* time-series of quantities measured from a dynamical system may be sufficient to construct a model

that preserves its essential structure. Takens (1981) developed and clarified the mathematical evidence for this proposal. This was considerably generalized by Sauer, Yorke & Casdagli (1991), and more recently Stark, Broomhead, Davies & Huke (1997) have extended these results to the more general case of stochastic dynamical systems.

Sauer et al. (1991) have suggested that the foundations of these ideas are to be found in differential topology. For example, a seminal theorem in this field (Whitney 1936) is that any m-dimensional manifold M can be mapped by a diffeomorphism[2] into Euclidean space $\mathbb{R}^d$ if $d > 2m+1$. Moreover, the subset of all possible smooth maps from M to $\mathbb{R}^{2m+1}$ that are also diffeomorphisms is both open and dense in the function space. As Sauer et al. (1991) point out a single measurement of a dynamical system is a map from the system's state to the real line. As such, the significance of Whitney's result is that *almost every*[3] set of $2m+1$ independent measurements of a dynamical taken simultaneously is sufficent to reconstruct the dynamical system in the measurement-space. The manifold M and its vector-field $\dot{x}$ are *embedded* in the measurement-space.

The more recent result by Takens (1981) may be understood in in terms of this embedding theorem. Takens considers the case of a dynamical system $f(\vec{x}, \delta) : M \mapsto M$ and the *delay-coordinate* map, $D : M \mapsto \mathbb{R}^{2m+1}$. This map D is defined as simply a time-series of scalar measurements $z = \{y_t, y_{t+1}, \ldots, y_{t+2m}\}$ obtained from this system, where $y = g(\vec{x})$. It is clear that

$$z = \{y_t, y_{t+1}, \ldots, y_{t+2m}\} = \{g(\vec{x}_t), g \circ f(\vec{x}_t), \ldots, g \circ f^{2m}(\vec{x}_t),\}$$

where f^n is the composition of f n-times. In other words, the sequence z of $2m+1$ measurements $y = g(\vec{x})$ is in fact a function of a single point or state $\vec{x}$ of the hidden dynamical system. The *delay-coordinate* map D maps each state $\vec{x}$ of the hidden system to a point in $\mathbb{R}^{2m+1}$. Takens (1981) demonstrated that with minimal assumptions about the hidden dynamical system[4] , the set of *delay-coordinate* maps D that are also diffeomorphisms is both open and dense in the space of maps D. In *almost every* case, the hidden dynamical system is *embedded* within the delay-coordinate measurement space.

Sauer et al. (1991) have considerably elaborated the Takens (1981) embedding theorem. They define both a delay coordinate map $D' : M \mapsto \mathbb{R}^s$, where s is an integer arbitrarily greater than $2m+1$, and a smooth transformation of this map, $\phi : D' \mapsto \mathbb{R}^{2m+1}$. In the spirit of Takens (1981), Sauer et al. (1991) demonstrate that the set of these composite functions $\phi \circ D' : M \mapsto \mathbb{R}^{2m+1}$ that are also diffeomorphism is open and dense in the function space.

The theorems of Takens (1981) and Sauer et al. (1991) apply to deterministic dynamical systems. These are systems whose entire future evolution can be be determined from precise knowledge of the system's state. As real world systems are inevitably coupled with sources of external noise, the generality of these theorems may seem limited. Stark et al. (1997) have shown, however, that the embedding theorems can be generalized to a much less restricted class of stochastic dynamical systems. They consider a discrete time system where at each time step one of k different discrete-time maps $f_\omega : M \mapsto M$ is chosen, where $\omega = 1, \ldots, k$. As in Takens (1981), they define the *delay-coordinate* map, $D : M \mapsto \mathbb{R}^{2m+1}$ and show that in the stochastic systems under consideration the set of maps D that are also diffeomorphism is open and dense in the function space.

State-Space reconstruction in neural systems

While these results are obviously important for the general problem of nonlinear time-series analysis, their relevance for the problem of language learning may be limited. The problem of language learning does not fit neatly into the scenarios considered by Takens (1981), Sauer et al. (1991) and Stark et al. (1997). This is primarily due to the fact that the output of the language generating dynamical system is a sequence of symbols rather than a real-valued scalar. In addition, the stochastic system considered by Stark et al. (1997) might not be general enough to describe the arbitrary stochastic dynamical system that is here taken to be the language generator. More importantly, these theorems consider and explain certain *sufficient* conditions and do not lead naturally to a general algorithmic procedure for reconstructing state-space. For example, Taken's theorem demonstrates that the coordinate space of $2m+1$ scalar measurements is sufficient to embed the generating dynamical system of dimensionality m. Practically, however, this just means that the coordinate space of a *finite* number of scalar measurements is sufficient for the embedding. It does not indicate how it can be known that an embedding has in fact occurred. What is necessary, therefore, is an *objective function* that may be optimized to produce a reconstruction of the dynamical systems from its outputs.

Crutchfield & Young (1989) introduce ε-machines as a general procedure for state-space reconstruction. They propose that the state of the ε-machine uniquely corresponds to the state of a dynamical system emiting a symbol sequence if it can be shown that its state renders the future of the symbol sequence conditionally independent of its past. In other words, if the probability distribu-

[2] A *diffeomorphism* from M to N is a one to one map, where the map and its inverse are diferentiable.

[3] The fact that the set of maps that are also diffeomorphisms is an *open* subset of the function space means that any arbitrarily small perturbation of a diffeomorphism is also a diffeomorphism. The fact that the set is *dense* means that every point in the function space is arbitrarily close to a diffeomorphism. In addition, Sauer et al. (1991) have shown that *almost every* map in the function space is a diffeomorphism, in that the complement to this subset is of measure zero. In other words, the likelihood of an arbitrary map also being a diffeomorphism is *probability one*, or infinitely likely.

[4] In particular, it is assumed that the dynamical does not contain periodic orbits that are exactly equal to (or exactly twice) the sampling rate of the measurement function $y = g(\vec{x})$.

tion over future sequences of symbols is independent of the past symbols *given* the state of the ε-machine, then the ε-machine uniquely labels the state of the dynamical system generating the symbols. The ε-machine can then be taken as a model of the unseen symbol generating dynamical system[5].

On the basis of the embedding theorems and the ε-machine of Crutchfield & Young (1989), an objective function for state-space reconstruction may be introduced. The objective is to *model* the dynamical system $f(\vec{x},\delta)\colon M \mapsto M$, and this can be defined as learning a structure-preserving map from the manifold M to a second topological *model-space* N. If the probability distribution over sequences of symbols emitted by the dynamical system defined on M is independent of its past symbols *given* the state of the *model-space* N, then N smoothly and uniquely labels the state of the dynamical system generating the symbols. The trajectory of states on N can then be taken as a model of the unseen symbol generating dynamical system N. This idea may be illustrated by means of a neural system.

A system of cortical neurons can be minimally modeled by a set of n coupled nonlinear differential equations,

$$\dot{y}_i = -y_i + \sum_{j=1}^{j=n} w_{ij}\sigma(y_i) + I_i,$$

where σ is a smooth and monotonic transfer function, y_i is the soma potential of neuron i, resulting from a weighted sum of its inhibitory and excitatory inputs. I is the external input to the system. Clearly, this system is a dynamical system defined on a n-dimensional manifold N. In addition, the state of this system $\vec{y}_t$ at a given time t is a function of both its present input I_t and, through the action of its recurrent synapses, the history of previous input, $\{I_{t_0},\ldots,I_t\}$. In other words, the system's state at any given time is a smooth function of an entire sequence of inputs. This can be represented by the correspondence $\vec{y}_t = \psi(I_{t_0},\ldots,I_t)$. If the sequence of inputs $\{I_{t_0},\ldots,I_t\}$ represents the output $I_t = g(\vec{x}_t)$ of dynamical system $f(\vec{x},\delta)$ then it is clear that

$$\vec{y}_t = \psi(I_{t_0},\ldots,I_t) = \psi\left(g(\vec{x_{t_0}}), g\circ f(\vec{x_{t_0}}),\ldots,g\circ f^t(\vec{x_{t_0}})\right),$$

where f^t is the composition of f t times. The state $\vec{y}$ of the neural system is a function of the state $\vec{x}$ of the hidden dynamical system.

The neural system $\dot{y}_i$ on N is a diffeomorphism of the dynamical system $\dot{x}$ on M, if the state $\vec{y}$ smoothly and uniquely labels the state $\vec{x}$. If the future inputs to the neural system are stochastically independent of the past inputs, *given* the state $\vec{y}$ of the system then N and M are diffeomorphically equivalent. If the probability of the future inputs to the neural system, conditioned on its state $\vec{y}$, is not further sharpened by acquiring information about the previous inputs to the system then there is a structure preserving map between the two systems.

Network simulations

In this paper, it is taken that a language (or a sequence of symbols) is produced by a continuous dynamical system. To learn this dynamical is to learn the statistical structure of the language. By hypothesis, this can be accomplished by embedding the hidden dynamical system in a second model space. To maximize prediction of future states given present ones is effectively to seek such an embedding. As such, it should be the case that if a recurrent neural network is trained on a corpus of natural language (in the now familiar style introduced initially by Elman (1990)) it should develop a state space that is a model of the generating process of the language. One manifestation of this would be that sentences, judged (by human observers) to be structurally similar, should also be clustered in the state space of the neural system.

To explore this hypothesis further, a simulation of an idealized neural system was performed by implementing the system of coupled equations,

$$\dot{y}_i = -y_i + \sum_j w_{ij}\sigma(y_i) + \theta_i + \sum_k w_{ik}I_k,$$

$$O_l = \sigma\left(\sum_i w_{li}(y_i\right),$$

$$\sigma(\zeta) = \left(1+e^{-\zeta}\right)^{-1},$$

where y_i is the state of the neuron and can be viewed as representing its mean soma potential, θ_i is a bias term and I_i is external input. O_l is the output of the system which "reads off" the recurrent network. There were 120 neurons in the recurrent network. The input was a 250 dimensional bit vector, described below. The output was likewise a 250 dimensional vector. For the purposes of computer simulation, a difference equation was used,

$$y_i^{t+\Delta t} = (1-\Delta t)\,y_i^t + \Delta t\sum_j w_{ij}\sigma(y_i) + \Delta t\,\theta_i^t + \Delta t\sum_k w_{ik}I_k^t,$$

[5] In a dynamical system, the entire evolution of the system is described by its trajectory from $t_{-\infty}$ through t_0 to t_∞. The future trajectories of the system are conditionally independent of the past, *given* the present state of the system. In the ideal case of a deterministic and autonomous system, the future trajectory of the system, $X[t_0,t_\infty)$, can in principle be determined from the present state of the system, $X(t_0)$. Absolute knowledge of the system's state X at t_0 provides absolute knowledge of the future trajectory $X[t_0,t_\infty)$. No information about the system's prior trajectory $X(t_{-\infty},t_0]$ is necessary. In a stochastic dynamical system (where, for example, at irregular points in time there is coin toss of an k-sided coin to choose between k different set of differential equations), a similar situation occurs. While the future is not entirely predictable on the basis of the present state in this system, no *increase* in information about the future is gained by knowing the past. In other words, the future trajectory of the system is stochastically independent of the past, *given* the state of the system. The case of a stochastic system can be seen to generalize to the case of a dynamical system driven by external input.

This was obtained by an approximation of its continuous counterpart. Δt was a variable parameter which could be manipulated for finer approximations of the underlying continuous system.

The data-set used for network learning was a corpus of natural language amounting to over 10 million words. The corpus comprised 14,000 documents, the average length of each document being approximately 700 words. All documents were in a plain-text and untagged format. They were obtained from publicly available electronic text archives on the internet[6]. No explicit criteria were used when selecting documents other than that cover a wide range of subject matters such as science, social science, literature, children's stories, history, law and politics.

Altogether, the entire corpus contained a vocabulary of 115,000 words. Of these, a set of 50,000 accounted for over 99% of the total number of words in the corpus. Only the members of this set were used for training the network, the infrequent words having been deleted. Each of these 50,000 words was coded by being randomly assigned to a unique bit vector of 247 zeros and 3 ones (there are over 2.5 million possible combinations to choose from). While this random coding scheme introduced some spurious correlations between words, the average correlation between words was close to zero[7].

The network was presented with the entire corpus as a sequence of words, one word at a time. The network was trained to predict its future word-input given its present word-input. The synaptic weight parameters were adapted using the continuous version of back-propagation through time due to Pearlmutter (1989). In this procedure, the minimum of the cross-entropy objective function was sought by calculating the derivatives of this function with respect to each weight parameter at each time "tick" Δt of the 50 previous time steps.

With a learning rate parameter of .01, and a Δt parameter of .25, the network was trained for 46 passes through the corpus. At this time, the learning rate parameter was annealed to .001, and the Δt parameter was lowered to .1. Training was continued for another five passes through the entire corpus. The performance of the network at predicting future words could be adequately assessed using the a method of ratios between squared errors,

$$R_i = \frac{\sum_t \left(d_i^t - y_i^t\right)^2}{\sum_t \left(d_i^t - d_i^{t-1}\right)^2},$$

where d_i^t is the target or to-be-predicted outcome for neuron i at time t. The denominator of this ratio specifies the sum squared differences between the target outcome and the target at the previous step. This ratio is useful as the best prediction a random-walk model can make would be to predict the same value for the future as is obtained at the present. Thus, if the ratio is greater than 1.0 the network is performing worse than a chance model. At values less the 1.0, the network is performing better than a chance model. A value approaching 0, would indicate perfect predictive accuracy.

On the final pass through the corpus, the mean performance ratio for the training data was .4767. Furthermore, a validation set which comprised 1000 unseen documents was prepared. The mean performance ratio on this set was .4989. These values indicate substantial predictive performance and generalization abilities by the network. They compare very favorably to mean performance ratios usually obtained in non-linear time series prediction tasks (Weigend & Gershenfeld 1993).

Discriminant function analysis

If a neural network learns the statistical structure of the language , its state space should have topological organization based on a similarity principle. For example, sentences that are similar in content should cluster in compact neighborhoods of the state space. An ideal experimental test of this would be to have reliable human judges classify a large set of sentences on the basis of their content, and then to compare this with a network's classification of the same set of sentences. To the extent that the network's classifications are close to those of human judges, the network would have met a behavioral criterion for language comprehension.

To adequately assess generalization abilities, a large set of sentences would be required. Such an experiment would be laborious to conduct. Fortunately, however, data-sets of labeled or categorized documents (rather than sentences) are readily obtainable, as these are regularly used as benchmark tests of text categorization techniques. In the experiment conducted here, sentences were extracted from labeled documents. Sentences were then assigned to the semantic class of the document from which they came. For example, sentences taken from a document assigned to the class 'motorcycling' would themselves be assigned to the semantic class 'motorcycling'. In this way, a large set of sentences could assigned a plausible, although somewhat limited, interpretation. The data-sets were the Reuters-21578 newswire data-set, the 20 newsgroups data-set[8] , and then a third set which was compiled for the purpose of this experiment from 6000 documents obtained from the Library of Congress, which had been previously classified by their Dewey Decimal categories

An appropriate test of the network's representational capacities would be to assess the probability that a sentence from a given semantic class would be assigned correctly to that class. To do this a linear discriminant function was used to divide the state space into (simply connected and convex) sub-regions based on semantic class. The discriminant function is a straightforward

[6]The main sources of the electronic texts were, Project Gutenberg, the Etext Archives, and archives.org.

[7]A more valid distributed code based the actual orthography of English words has been used by the author in previous simulations, but these will be reported here.

[8]The two data sets are available on the internet. See http://www.cs.cmu.edu/ textlearning and http://www.research.att.com/ lewis

Table 1: Accuracy of sentence classification.

Data Set	Accuracy
Library	83%
Reuters	75%
Newsgroups	69%
Mean	76%

linear transformation of the state space, such that the centroids of "training" sentences labeled by their class are made maximally distant from one another. The network's ability to categorize by semantic class can be assessed for a "test" set of sentences by assessing the probability that a given sentence from a certain semantic class would be correctly assigned to that class. The measure used was Mahalanobis distance. This measure is approximately proportional to an estimate of the posterior probability that a given sentence will correctly assigned to its appropriate class. 5000 sentences from each of the three data-set were used in this test. The results are illustrated in Table 1.

These accuracy rates are suitably high, and in fact compare favorably to state-of-the-art text categorization methods which use similar or identical data-sets (Nigam, Mccallum, Thrun & Mitchell 2000). It is reasonable to conclude from this that the state space of a recurrent neural network trained to predict word sequences becomes organized on basis on semantic similarity. Sentences and texts that are semantically similar are clustered into compact neighborhoods which can be discriminated by a simple linear function.

Conclusion

Temporal pattern recognition is not as theoretically sophisticated as its multidimensional and static counterpart. Here, an approach to temporal pattern learning is introduced that is based on recent results from dynamical systems theory. It is proposed that the reconstruction of system generating a language (or symbol sequence) is adequate for learning the statistical structure of temporal data. It is proposed that state-space reconstruction can be carried out in a straightforward manner in a recurrent neural network. Results showing pattern recognition of English sentences by the network are provided. These results are similar in kind to those obtained by Elman (1990) and in the many works that followed this paradigm. It is believed that the appropriate explanation of these now familar sets of results, is that the recurrent neural network has reconstructed the language generating process. Sentences that were produced by similar trajectories in the original systems are now modelled by similar trajectories in the recurrent neural network. It is clear, however, that this is not a definitive demonstration of state-space reconstruction and a more detailed analysis of temporal pattern learning using formal grammars is being currently undertaken (Andrews 2001).

References

Andrews, M. W. (2001), Language learning by state space reconstruction. Manuscript in preparation.

Bai-Lin, H. & Wei-Mou, Z. (1998), *Applied Symbolic Dynamics and Chaos*, World Scientific, Singapore.

Chomsky, N. (1963), Formal properties of grammars, *in* R. D. Luce, R. R. Bush & E. Galanter, eds, 'Handbook of mathematical psychology', Vol. 2, John Wiley and Sons, Inc., New York and London, pp. 323–418.

Crutchfield, J. P. & Young, K. (1989), 'Inferring statistical complexity', *Physical Review Letters* **63**(2), 105–108.

Elman, J. L. (1990), 'Finding structure in time', *Cognitive Science* **14**, 179–211.

Nigam, K., Mccallum, A., Thrun, S. & Mitchell, T. (2000), 'Text classification from labeled and unlabeled documents using em', *Machine Learning* **39**(2/3), 103–135.

Packard, N., Crutchfield, J. P., Farmer, J. & Shaw, R. (1980), 'Geometry from a time series', *Physical Review Letters* **45**(9), 712–716.

Pearlmutter, B. (1989), 'Learning state space trajectories in recurrent neural networks', *Neural Computation* **1**, 263–269.

Sauer, T., Yorke, J. A. & Casdagli, M. (1991), 'Embedology', *Journal of Statistical Physics* **65**(3–4), 579–616.

Stark, J., Broomhead, D. S., Davies, M. E. & Huke, J. (1997), 'Takens embedding theorem for forced and stochastic systems', *Nonlinear Analysis, Theory, Methods and Applications* **30**(8), 5303–5314.

Tabor, W. (1998), Dynamical automata, Technical report, Department of Computer Science, Cornell University.

Tabor, W. (2000), 'Fractal encoding of context-free grammars in connectionist networks', *Expert Systems* **17**(1), 41–56.

Takens, F. (1981), Detecting strange attractors in turbulence, *in* D. Rand & L.-S. Young, eds, 'Dynamical Systems and Turbulence', Springer-Verlag, Berlin and Heidelberg, pp. 366–381.

Weigend, A. S. & Gershenfeld, N. A., eds (1993), *Time series prediction: Forecasting the future and understanding the past*, Vol. 15, Addison-Wesley, Reading, MA 01867.

Whitney, H. (1936), 'Differentiable manifolds', *Annals of Mathematics* **37**(3), 645–680.

Comprehension of Action Sequences: The Case of Paper, Scissors, Rock

Patric Bach* (bach@mpipf-muenchen.mpg.de)
Günther Knoblich* (knoblich@mpipf-muenchen.mpg.de)
Angela D. Friederici° (orendi@cns.mpg.de)
Wolfgang Prinz* (prinz@mpipf-muenchen.mpg.de)

*Max Planck Insitute for Psychological Research; Amalienstr. 33, 80799 Munich, Germany
° Max Planck Institute for Cognitive Neuroscience, Stephanstr. 1a, 04303 Leipzig, Germany

Abstract

The basic idea of the present study is that it is useful to conceptualize the processes involved in action comprehension in a similar manner as the processes involved in sentence comprehension. One important question then is whether order and meaning of action sequences are processed sequentially or in parallel (analogous to syntactic and semantic processing in sentence comprehension). We conducted three experiments to address this issue. Participants were asked to detect violations of order, meaning, or both, in action sequences of the game Paper, Scissors, and Rock. The main results were that it took longer to detect single violations than double violations and that it was impossible to ignore any of the violations. This pattern of results suggests that the processes involved are highly automatic and that they are running in parallel.

Have you ever wondered why people in sports bars stare at the numerous TV-screens that are silently picturing sequences of actions like people in fancy dresses bumping forcefully into one another, or two people hitting a ball with their rackets? Obviously, in some situations observing other people's actions is more interesting than chatting. This is not only true for sports bars but also for French street cafés and many other places. But how do we understand what happens while we watch other people acting?

Whereas cognitive science has been making much progress regarding the processes involved in language comprehension (Friederici, Steinhauer, & Frisch, 1999; Garrod & Pickering, 1999; MacDonald et al., 1994; Rayner & Pollatsek, 1989), action comprehension has been hardly addressed. Although Schank and Abelson's (1977) script theory seems relevant, it mainly deals with the comprehension of extended texts and not action comprehension as in the examples above. Another well-known approach, the grammar of action (Goodnow & Levine, 1973), addresses only the construction of complex movements from simpler units in writing. Our research is guided by the idea that it might be useful to conceptualize the processes involved in action comprehension in a similar fashion as the processes involved in sentence comprehension.

Sentence and Action Comprehension

In order to understand a sentence it is necessary to parse it according to the grammatical rules that governed its formation and to understand the meaning of the words given the context of the sentence. Likewise, to understand an action sequence it is often necessary to parse the sequence according to rules that constrain the order in which one action follows another and to understand the meaning of single actions given the context of the sequence. Although there might be action sequences the order of which is hardly constrained, most of them follow rules that can be spelled out clearly.

The main question for action comprehension (as for sentence comprehension) then becomes how processing of the order of action sequences (in analogy to syntactic processing) and the processing of its meaning are related. There are essentially four possibilities: (1) The analysis of order generally precedes and influences the analysis of meaning. (2) The analysis of meaning generally precedes and influences the analysis of order. (3) The two processes run in parallel and do not affect each other. (4) They run in parallel and they do affect each other. Controversies in sentence comprehension have mainly focused on the possibilities that are analogous to 1 and 4. The garden path model (Frazier & Rainer, 1982) suggests that meaning does not influence the selection of the initial syntactic structure (syntax first). The connectionist theory of MacDonald et al. (1994) states that syntactical and semantic constraints are narrowing the possible interpretations of a sentence in parallel and interactively. However, (2) and (3) are also viable possibilities for action comprehension. For instance, in favor of (2) one could say that the meaning of a single action of a peer might be most relevant for organisms (for instance because they can be dangerous or not) and that therefore the action comprehension system analyses meaning before order. The experiments that follow focused mainly on the question of whether the order and meaning of an action sequence are analyzed in parallel or whether there is any sequential order of the two processes.

Experimental Paradigm

In our experiments, we used a simple action domain, the game Paper, Scissors, and Rock. In this game, the order of consecutive actions is well defined and only a small set of gestures is meaningful in its context. Moreover, the game is well known (at least in Germany), so that people are quite familiar with the rules. The game has a fixed sequence (see Figure 1): In the beginning, the two players hold their hands close to the chest (upper position) and form a fist. In the next step, they drop their fists to a position near their waists (lower position) synchronously, and lift them again. This pattern is repeated once. During the third downward movement both players form one of three gestures: paper, scissors, or rock. The winner is determined by the following rules: Rock dulls scissors (rock wins), scissors cut paper (scissors win), paper wraps rock (paper wins). There are three meaningful gestures, that is, paper, scissors, and rock. Other gestures produce violations of context, as for instance the thumbs up gesture. It is also easy to see what a violation of order might look like. There is only one valid structure that is defined by alternating hand positions:

Upper, lower, upper, lower, upper, lower.

Therefore violations of order can be introduced by changing the sequence in the following way:

Upper, lower, upper, lower, upper, *upper.*

Obviously, violations of context and order are independent of one another. A violation of order might occur, although a meaningful gesture is formed (e.g. paper in the upper position, see leftmost picture in Figure 2). Alternatively, a meaningless gesture can be formed without violating order (e.g. thumbs up at the lower position (see picture in the middle of Figure 2). In addition, both violations can occur at the same time as when displaying a meaningless gesture at the wrong position (e.g. thumps up at the upper position, see rightmost picture in Figure 2). By comparing reaction times for different types of violations one can determine how order and meaning are analyzed in this type of action sequence.

Experiment 1

The aim of Experiment 1 was twofold. The first aim was to determine whether it is easier to detect violations of order than violations of context. If processing the order of an action sequence is similar to syntactic processing, it might be completed before the meaning of an action is fully retrieved or it might even be a necessary condition for the analysis of meaning. Hence people might be faster in detecting violations of order than violations of context. If, on the other hand, the processing of meaning is prevalent in action comprehension, violations of context should be faster detected.

The second aim was to determine whether order and meaning are processed in parallel. Faster detection of either violation does not necessarily imply that sequence and meaning are analyzed in a serial fashion. Rather, the differences could be due to different processing times taken by two parallel processes. If order and meaning are processed in parallel it should take less time to detect double violations than either single violation. The rationale behind this assumption is that the process that is completed first will trigger the reaction indicating that the violation was detected. Alternatively, evidence accumulated by each process might converge to trigger the wrong reaction. If order and meaning of action sequences are processed sequentially, detecting a double violation should take the same time as detecting the single violation (of either order or context) that is detected fastest.

One problem with measuring RTs for the detection of order and context violations in action sequences is that the onset of the violation is not well defined. When watching a video film displaying a person playing paper, scissors, and stone, the syntactic violation may be detected as soon as the expected downward movement does not occur. The semantic error can only be detected as soon as a gesture is formed. To avoid these problems, we used still frames that displayed the upper and lower position of the hand three times, respectively. The changing frames create the impression of a movement and they are readily interpreted as action sequences (Stürmer, Aschersleben, & Prinz, 2000). With still frames, syntactic, semantic, and double violations have the same onset, that is, the onset of the frame that completes the sequence at the lower position.

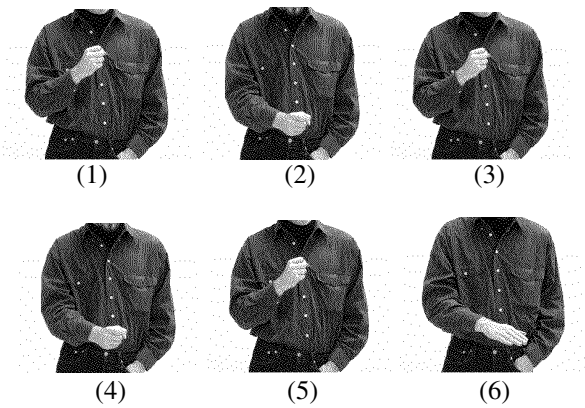


Figure 1: Sample sequence of pictures (correct).

Method

Participants. Sixteen participants, all students at the University of Munich, took part in the experiment, 9 of them male. They ranged in age from 24 to 35 years. The key assignment was counterbalanced across participants.

(a) v. of order (b) v. of context (c) double violation

Figure 2: Sample for last frame in action sequences with violation of order, meaning, or both.

Material. The material consisted of action sequences showing still frames of a person playing paper, scissors, and rock. There were six pictures displaying the actor forming legal gestures (Rock, Scissors, Paper), and six displaying him forming illegal gestures. The first five pictures of a sequence always showed a closed fist alternating between the upper and lower position. The sixth picture displayed one of the six possible gestures in the upper or lower position (see Figure 1 for a complete legal sequence). Hence, there were sequences ending with (1) a legal gesture at the lower position (correct), (2) an illegal gesture at the lower position (violation of context), (3) a legal gesture at the upper position (violation of order), and (4) an illegal gesture at the upper position. Figure 2 shows examples for the last frame in the violation conditions.

Procedure and Design. The experiment consisted of two blocks of 144 trials each. The order of action sequences was randomized in each block. To avoid response bias, correct action sequences were displayed as often as incorrect sequences. Hence there were 72 correct action sequences in each block, and 24 containing a violation of order, context, or both, respectively. The course of each trial was as follows: The first five frames were displayed for 500 ms, respectively. The sixth, critical frame was displayed for 1000 ms. The reaction time interval started with the onset of the last frame and lasted as long as the gesture was displayed (1000 ms). The participants judged whether there was a violation or not and pressed a left or a right key, accordingly. If they committed an error or did not react within the given time interval of 1000 ms, they were given error feedback.

Results and Discussion

Figure 3 displays the RTs for the detection of different types of violations during the first and second block of trials. Double violations (M = 540, S = 56) were faster detected than either violations of order (M = 569, S = *71*) or violations of context (M = 602, S = 57). Moreover, violations of order were faster detected than violation of context. These differences were present in both blocks, although RTs were generally faster in Block 2 (M = 559 ms, S = 66) than in Block 1 (M = 580 ms, S = 55). The mean RTs for responses to correct action sequences were 566 ms (S = 58). They were not included in the statistical analysis because they are hard to compare to RTs for violations because of the different base rate and the different type of response.

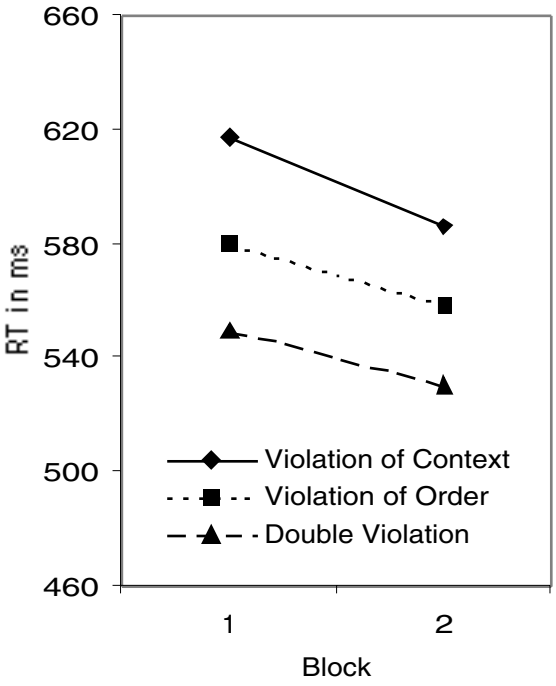


Figure 3: RTs for detecting different types of violations.

The RTs for the three error conditions were entered into a 3 x 2 repeated measurement ANOVA with the factors Type of Violation (Order, Context, vs. Double) and Block (1st vs 2nd). The analysis revealed significant main effects for Block, F (1, 15) = 9.9, p < .01, and Type of Violation, F(2, 30) = 28.4, p < .001, but no significant interaction. The RTs differed significantly for different types of violations. Duncan tests showed that double violations were detected significantly faster than either violations of order or context (both p < .001). In addition, there was also a significant difference between the two types of single violations (p < .001).

The pattern of RTs is in favor of a parallel processing explanation. Double violations were detected faster and more reliably than either violations of sequence or context. This is in favor of the claim that either of two

parallel processes being completed first can trigger the reaction indicating that a violation was detected. Order violations were faster detected than context violations right from start. Hence, it seems that analyzing the order of a sequence is completed before the meaning of an action is fully processed. An alternative explanation for the different speed with which violations of order and context can be detected is perceptual saliency. Violations of order might have been easier to detect because a wrong hand position is perceptually more salient than a hand forming a different gesture.

Experiment 2

The second experiment was conducted to rule out the perceptual saliency explanation. To do so, we replicated Experiment 1 using a set of stimuli in which the perceptual saliency of order and context violations was more comparable. The deep structure of the action sequence, five alternations of the same gesture and formation of the target gesture with the last alternation, did not change. However, the alternations were now expressed as hand turnings instead of arm movements (see Figure 4). Hence both, violations of order and context now depended on the configuration of hand and fingers. Violation of order consisted in a missing turn of the hand, violations of context consisted in invalid gestures, as in Experiment 1 (see Figure 5).

If higher perceptual saliency caused the faster detection times for violations of order in Experiment 1 the detection times should be slower or equally fast as for violations of context under the new conditions. If faster processing of sequence information caused the difference the results should be the same as in Experiment 1. It is quite likely that such a difference would not occur right from start because people have to link the new type of alternation (hand turning) to the familiar deep structure of the sequence before optimal detection of order violations becomes possible. Hence, in order to allow this link to be established we added a further block of trials. For double violations we expected the same pattern as in Experiment 1, that is, faster detection than for either single violation.

Method

Participants. Fifteen participants, all students at the University of Munich, took part in the experiment, five of them male. They ranged in age from 24 to 35 years.

Material. A different set of stimuli was used. The fixed pattern of alternations consisted of hand turnings while the hand position remained fixed. The critical manipulation occurred with the last frame. The gesture displayed could either be correct, a violation of order, context, or both (see Figure 5 for examples). There were three different valid gestures and three that violated the context. There were two versions of each gesture, one displaying the face of the hand (correct) and one displaying the backside (violation of order).

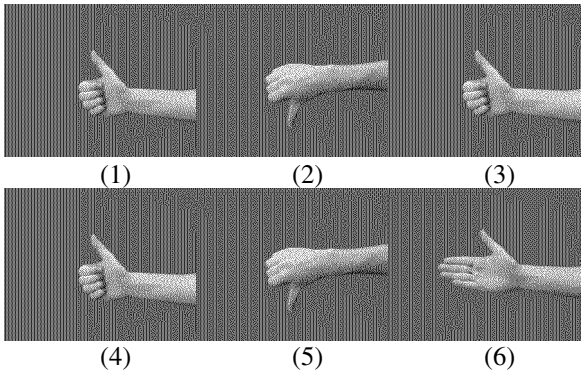


Figure 4: Sample sequence of pictures (correct).

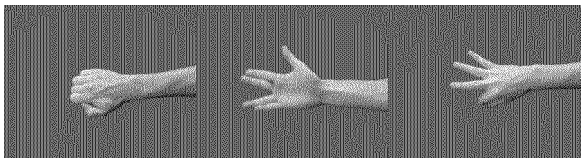


Figure 5: Examples for violation of order (left: wrong orientation), context (middle), and double violation (right) in Experiment 2.

Procedure and Design. The procedure was the same as in Experiment 1 with one exception. A third block of 144 trials was added.

Results and Discussion

Participants pressed the wrong button in about 3% of the cases and reacted too late in about 2% of the cases. Figure 3 displays the RTs for the detection of different types of violations during the first, second, and third block of trials. Overall, it took participants longer to detect violations of context (M = 615, S = 64) than violations of order (M = 596, S = 76). However, this difference occurred only after the first block. As expected, double violations were faster detected than either single violation right from start (M = 561, S = 70). Participants became faster during later blocks. Mean RTs were 617 (S = 56), 584 (S = 63), and 572 (S = 70) during Block 1, 2, and 3.

The RTs of the three error conditions were entered into a 3 x 3 repeated measurement ANOVA with the factors Type of Violation (Order, Context, vs. Double) and Block (1, 2, vs. 3). There were significant main effects for Type of Violation, $F(2, 28) = 14.7$, $p < .001$, and Block, $F(2, 28) = 20.9$, $p < .001$, the interaction between both factors was only marginally significant, $F(4, 56) = 2.0$, $p = 0.1$. Duncan-tests showed that the RTs for violation of order were significantly faster than for

violations of context in Block 2 and 3, but not in Block 1. The difference between double violations and either single violation was significant in each Block.

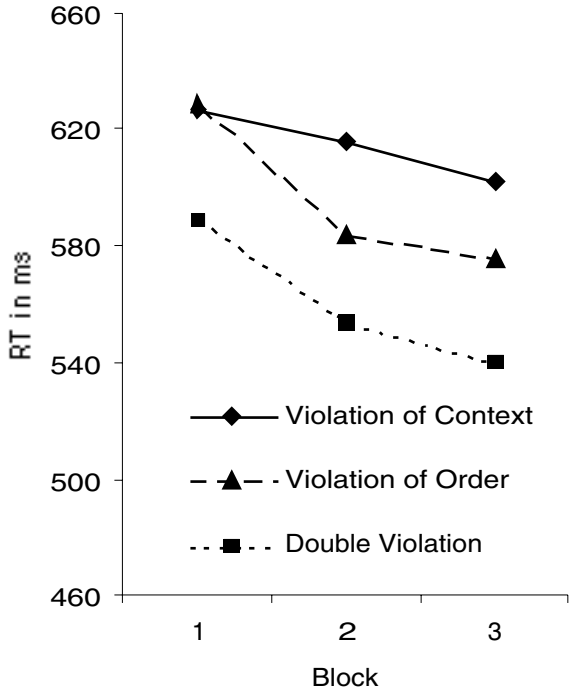


Figure 6: RTs for detecting different types of violations.

The results basically replicate Experiment 1. After Block 1, the RTs were almost numerically identical to those observed in the first experiment. This suggests that it was not perceptual saliency that created the RT differences between violations of order and context. Rather it seems that sequence information is faster processed, at least for simple sequences as in paper, scissors, and rock. The lack of difference during the first block is probably due to the necessity of establishing a link between the new perceptual cues and the familiar deep structure. The finding that double violations were faster detected right from start further supports the claim that order and meaning of action sequences are processed in parallel.

Experiment 3

It has often been claimed that parallel processes do not require attention and that they are highly automatic. Experiment 3 was conducted to determine whether this is also true for the processes involved in analyzing the order and meaning of action sequences. We used an interference paradigm to provide evidence for the automatic nature of the processes involved. The first block of Experiment 3 was identical to that of Experiment 2. In the following blocks we asked participants to ignore violations of order or context while detecting the other type of violation, respectively. In a first block, they were asked to detect violations of order and to ignore violations of context, that is, sequences containing the latter violation should be judged as correct. In the other block they were asked to detect violations of context and to ignore violations of order, that is, sequences containing a violation of order should be judged as correct.

If order and meaning of action sequences are analyzed automatically the violations to be ignored should interfere with judging a sequence as correct. If the processes are not automatic there should be no differences in RTs between action sequences containing an interfering violation and truly correct action sequences. Moreover, if sequence processing is faster than the processing of meaning a violation of order should interfere more strongly with the correct response than a violation of context.

Method

Participants. 20 participants, all students at the University of Munich, took part in the experiment, 8 of them male. They ranged in age from 24 to 35 years. The order of blocks was counterbalanced.

Material. The material used was identical to that of the second experiment.

Procedure and Design. The experiment consisted of three blocks of 144 trials each. The first block was the same as in Experiment 2. At the beginning of the second and third block participants received an instruction to attend to one of the single violations only and to ignore the other (e.g. detect violation of context/ignore violation of order). During the third block participants attended to the other type of violation and ignored the other (e.g. detect violation of order/ignore violation of context). The order of blocks 2 and 3 was counterbalanced across participants.

Results and Discussion

Participants pressed the wrong button in about 1% of the cases and reacted too late in about 3% of the cases. The RTs for different types of violations during the first block were indistinguishable from Experiment 2. Because of the space constraints, we will report only RTs for trials in which action sequences should be judged as correct. Figure 7 shows the results.

RTs were generally faster for correct sequences (M = 522, S = 59) than for sequences that contained an interfering violation (M = 580, S = 82). The difference was larger when a violation of order interfered (82 ms) than when a violation of context interfered (24 ms).

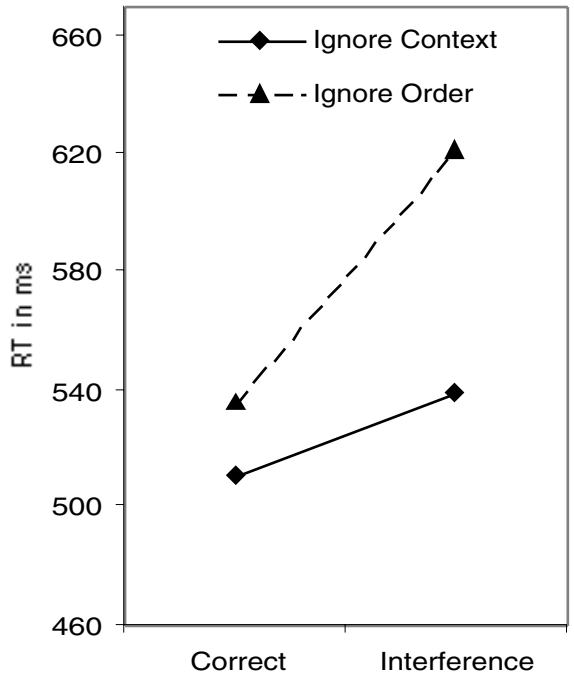


Figure 7: RTs with and without interfering violation.

The RTs were entered into a 2 x 2 repeated measurement ANOVA with the factors Type of Sequence (Correct vs. Interfering Violation) and Type of Interference (Order vs. Context). There were significant main effects for the factors Type of Sequence $F(1, 19) = 27.5$, $p < .001$ and Type of Interference $F(1, 19) = 86.6$, $p < .001$, as well as an interaction between the two factors $F(1, 19) = 23.1$, $p < .001$.

The pattern of results is consistent with the claim that the processes involved in analyzing order and meaning of action sequences are highly automatic. Otherwise one would not expect the huge interference effects that were observed in Experiment 3. It seems impossible to focus on one or the other aspect of an action sequence and to ignore the other. The stronger effect in the condition with interfering violations of order further supports the claim that analyzing the order of an action sequence takes less time than analyzing it's meaning.

General Discussion

The general pattern of results suggests that in action comprehension the analysis of order and meaning proceeds in parallel. Moreover, both processes seem to be highly automatic, that is, none of the two aspects of action sequences can be easily ignored. A further result is that analyzing the order of a sequence might be completed before its meaning is fully analyzed. However, further experiments are needed to determine whether the faster processing of order is a general phenomenon or whether it depends on the complexity of the task at hand.

The possibility to conceptualize action comprehension in an analogous manner to language comprehension does not necessarily imply that the cognitive processes involved in both domains are the same or that they are governed by the same principles. However, this strategy provides an opportunity to relate findings from both domains and to detect interesting parallels or differences. Currently, we are conducting ERP studies to determine whether violations of order and context in action sequences evoke the same ERP components as syntactic and semantic violations in spoken sentences (Friederici et al., 1999). The results of such studies might provide more conclusive answers to the question of whether there are cognitive processes which are actually contributing to both, action and language comprehension (Rizzolatti & Arbib, 1998).

References

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology, 14*, 178-210.

Friederici, A. D., Steinhauer, K., & Frisch, S. (1999). Lexical integration: Sequential effects of syntactic and semantic information. *Memory & Cognition, 27*, 438-453.

Garrod, S., & Pickering M. J. (1999). *Language Processing*. Hove: Psychology Press.

Goodnow, J. J., Levine, R. A. (1973) "The grammar of action": Sequence and syntax in children's copying. *Cognitive Psychology. 4*, 83-98.

MacDonald, M. C., Pearlmutter, N. J., Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review. 101*, 676-703

Rayner, K., Pollatsek, A. (1989). *The psychology of reading*. Englewood Cliffs,: Prentice-Hall.

Rizzolatti, G., & Arbib, M. (1998). Language within our grasp. *Trends in Neurosciences, 21,* 188-194.

Schank, R. C., Abelson, R. P. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures.* Hillsdale: Lawrence Erlbaum.

Stürmer, B., Aschersleben, G., & Prinz, W. (2000). Correspondence effects with manual gestures and postures: a study on imitation. *Journal of Experimental Psychology: Human Perception & Performance*, 26, 1746-1759.

Toward a Model of Learning Data Representations

Ryan Shaun Baker (rsbaker@cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Albert T. Corbett (corbett+@cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Kenneth R. Koedinger (koedinger@cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Abstract

The use of graphs to represent and reason about data is of growing importance in pre-high school mathematics curricula. This study examines middle school students' skills in reasoning about three graphical representations: histograms, scatterplots and stem-and-leaf plots. Students were asked to interpret graphs, select an appropriate graph type to represent a relationship and to generate graphs. Accuracy levels varied substantially across the three tasks and three graph types. The overall pattern of results is largely explained by the varying ease of transfer of student knowledge from a simpler graph type, based on surface similarity.

Introduction

External graphical representations are of considerable importance in problem solving. Considerable research has taken place over the last two decades on the different mechanisms through which graphical representations assist their users in drawing inferences (Larkin & Simon, 1987; Stenning, Cox, and Oberlander 1989).

In this paper we take up the use of representations at a very early point – at the point when a student is just learning to generate and interpret a representation – and ask what some of the major challenges are in learning these skills. There has been growing interest in attempting to teach these skills to students as young as those in the third through eighth grades[1] (NCTM 2000), but there is considerable evidence as well that these skills have not yet been developed by many undergraduates (Tabachneck, Leonardo, and Simon 1994).

We take up this subject in the context of developing a cognitive model of how novices generate and interpret some of the simpler representations used in data analysis. This model is designed with production-rule logic, in ACT-R (Anderson 1993). In this process, we hope to follow in the footsteps of some of the successful cognitive models of novices developed in other domains such as algebra problem solving (Koedinger & MacLaren 1997).

One area which might considerably influence students' performance on these tasks is transfer of the knowledge students already have of generating and interpreting other representations. Since students are taught different sets of representations at different grade levels (NCTM 2000), it is quite plausible that an important model for learning new representations will be the representations encountered earlier. Previous research into when transfer occurs shows that transfer can happen between exercises taking place in different representations, through mechanisms such as analogy, and that transfer can occur between similar processes (Novick 1988, Novick and Holyoak 1991, Singley and Anderson 1989). Hence, we seek to find out if and how these processes extend to the very first stages of learning how to use and generate a representation.

We are interested both in positive transfer, and in overgeneralization, where knowledge is transferred inappropriately. Scanlon's (1993) research in the use of representations for physics problem-solving provides some excellent examples of overgeneralization in the interpretation of different graphical representations. Additionally, other research has shown that misconceptions in physics, arising from overgeneralization of previously learned knowledge, causes long-term difficulties in correctly learning new material. How best to deal with such misconceptions is an active question in the research literature, with some arguing for a curricular strategy which acknowledges the appropriate contexts for certain conceptions and helps students see when they are inappropriate (NRC 1999).

In this paper, we present results and analysis of a empirical study we conducted in this domain, investigating novice performance (with an eye towards transfer effects) on interpreting, generating, and selecting representations important to early data analysis.

Domain

Representations

This study focuses on three graphical representations of data: histograms, scatterplots, and stem-and-leaf plots.

A *histogram* depicts a frequency distribution, as displayed in Figure 1. A set of interval categories (as in Figure 1) are represented in the X axis, and the frequency of each category is represented by the height of the corresponding vertical bar. A *stem-and-leaf plot,* shown in Figure 2, also

[1] Between the ages of 7 and 13.

displays frequency distribution data – the frequency of occurrence in this case for values between 0 and 99. In Figure 2, a distribution of 30 values, ranging from 4 to 97, is displayed. The higher order "tens" digit of the values form 10 categories down the left side of the graph. The lower order "ones" digit of each observed value is displayed in an ordered row to the right of the associated tens digit. Finally, a *scatterplot* employs a Cartesian plane to represent the relationship between two quantitative variables, as displayed in Figure 3. Each axis represents one of the variables, and the points represent paired values of the variables.

These three representations were selected because they are featured in most middle school math curricula and to systematically vary graph characteristics. Note that histograms and stem-and-leaf plots each portray univariate frequency distributions, although their surface features are dissimilar. The stem-and-leaf plot looks more like a table, frequency is depicted horizontally rather than vertically, and the frequency count is not directly depicted. In contrast, histograms and scatterplots share some superficial similarities – each has two numerically labeled axes – but they represent very different types of information.

A fourth type of graph, which was not included in the experimental tests, will be relevant in interpreting student performance. This is a *bar graph*, as depicted in Figure 4. A bar graph displays the values of a categorical variable along its X axis, and of a related quantitative variable along its Y axis.

Teacher Predictions

In this study, students are asked to (a) interpret graphical representations, (b) select the appropriate representation for different types of data display, and (c) generate different representations. We asked the two teachers in our sample classes to predict how their students would perform. The teachers predicted that students would perform about equally well on interpretation and generation, and poorly on selection. They predicted that students would be most successful with histograms, next most successful with scatterplots (because scatterplots are more conceptually challenging) and would perform worst on stem-and-leaf plots (because they are the least familiar to students).

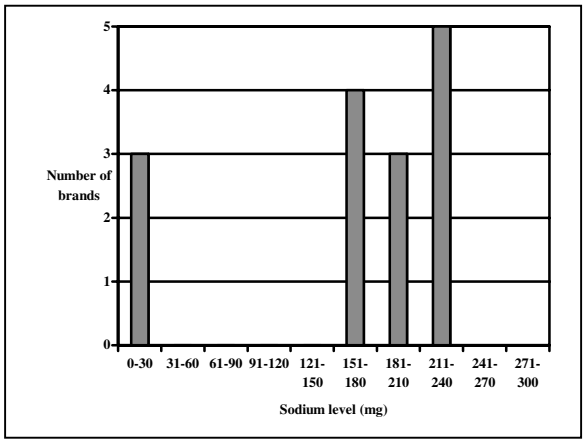


Figure 1: The histogram used in the interpretation exercises

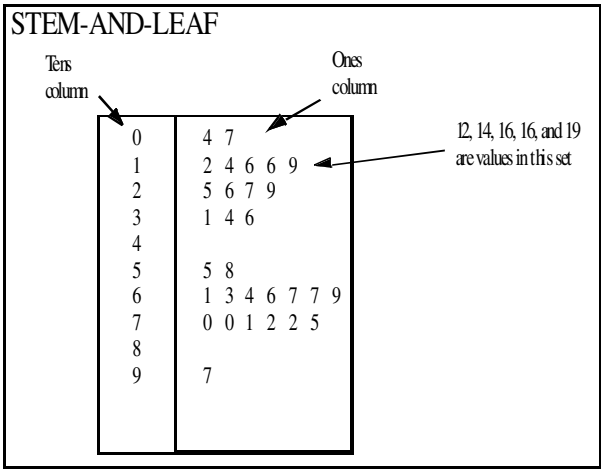


Figure 2: The drawing of a stem-and-leaf plot we used in our refresher sheet

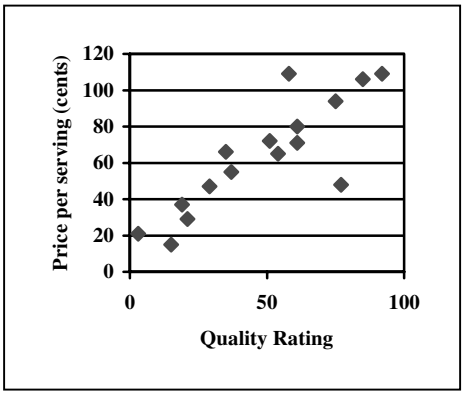


Figure 3: The scatterplot used in the interpretation exercises

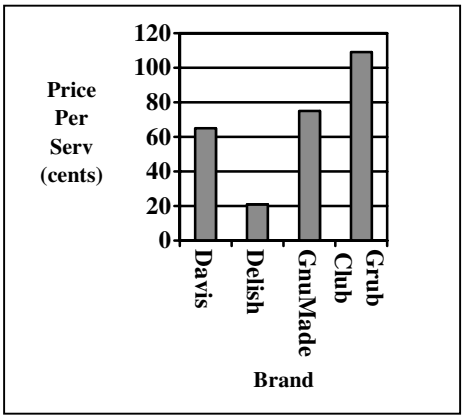


Figure 4: A Bar graph

Methods

Participants

The participants were 52 8th and 9th grade students from three mainstream pre-algebra classes in two Pittsburgh suburbs, half male and half female. The study was conducted prior to the year's data analysis unit; students had some exposure to histograms, scatterplots and stem-and-leaf plots in the last two years, and considerably greater exposure to bar graphs before that point.

Design

In the study, each of the participants completed 3-4 exercises in which they were asked to generate ("draw") a histogram, scatterplot, or stem-and-leaf plot, to answer a set of interpretation questions for one of those representations, or to select the most appropriate representation for a particular question. Four different forms were used, with questions on each form chosen such that neither the same category of task nor the same representation were assigned twice in one form. Within each form, the order of the exercises was rotated for different students to prevent order effects. These forms also included exercises involving box plots and tables, but we neither expected nor found the kind of effects we found for the representations discussed here.

The generation exercises gave the student a data set, in the form of a table, and asked them to draw the given representation of that data. The exercise statement read as follows:

> Please draw a [scatterplot, histogram, stem-and-leaf plot], showing all of the data in this table. Show all work. Feel free to use graph paper, if necessary.

The interpretation exercises gave the student a drawing of that representation and a set of questions to answer using that representation (see Figure 1and Figure 3).

The interpretation exercises had three types of questions, both multiple-choice and open-ended. The first type were straightforward questions typically asked for the target representation. These required no understanding of the representation's global properties (for the histogram, "How many brands of creamy peanut butter have between 0 and 30 mg of sodium?"). The second were also typical, but required understanding of the representation's global properties -- properties which require the student to make inferences (Stenning, Cox and Oberlander 1995, Leinhardt, Zaslovsky, and Stein 1990). "Is there a relationship between quality and price? Answer yes or no." is one such question for scatterplots. Finally, the third type were questions that are not typically asked for the target representation, but could be answered through productions more appropriate for another representation (for the scatterplot, "What is the price of the brand with a quality rating of 3?").

The representation selection questions gave the students a data set, in the form of a table, a question to answer (such as "What type of graph would be best for determining whether or not there is a relationship between price and quality?"), and four choices.

The students were also given a sheet with a picture of the four types of representations directly addressed in the exercises (histograms, scatterplots, stem-and-leaf plots, and box plots – bar graphs were not included in this sheet, nor mentioned in the study). An example from this sheet is shown in Figure 2. We did this in order to prevent the forgetting of terms from having an effect on the students' performance.

Scoring

For generation exercises, we developed rubrics for completely correct solutions (no features incorrect), and solutions that had the correct surface features (with the same general appearance as a correct solution – axes and bars or dots). For interpretation and representation choice exercises, we scored answers either completely right or wrong.

Results and Discussion

Performance accuracy in the graph interpretation, generation and interpretation tasks is displayed in Table 1. Students performed moderately well overall on graph interpretation, averaging 56% correct. However, there was large difference between performance on different representations – the 15 students interpreting histograms performed considerably better (average of 96% correct) than the 12 students interpreting scatterplots (average of 56% correct on the open-ended questions) ($t(25)=4.925, p<0.0001$). Both groups performed considerably better than the 13 students interpreting stem-and-leaf plots (average of 17% correct) (for scatterplots versus stem-and-leaf plots, $t(23)=4.109$, $p<0.001$; for histograms versus stem-and-leaf plots, $t(26)=12.191$, $p<0.0001$). In contrast, student performance on graph selection and graph generation was quite poor. Students were not better than chance accuracy (1 out of 4, 25%) in graph selection. Furthermore, they were completely unsuccessful at generating histograms and scatterplots. Performance by the 15 students who attempted to generate stem-and-leaf plots was relatively poor at 20% completely correct, but was marginally significantly better than the performance of the 12 students attempting to generate histograms (0% completely correct) and the 12 students attempting to generate scatterplots (0% completely correct), using a test of the significance of independent proportions. ($z=1.64$, $p<.11$).

The teachers accurately predicted that students would struggle with graph selection problems. Their predictions that histograms would be easiest and stem-and-leaf plots hardest corresponded with the data, but only for graph interpretation. Their expectation that students would have comparable success with generation and interpretation proved dramatically incorrect. Nathan and Koedinger (2000) report a similar result, that experienced teachers sometimes exhibit an "expert blindspot" and, in some cases,

Table 1: Average student performance in graph generation and interpretation. Percent which students would be expected to get right through guessing is placed in parentheses where appropriate.

	Histogram	Scatterplot	Stem-and -leaf plot
Generation	0%	0%	20%
Interpretation (open-ended questions)	95%	56%	17%
Selection	15% (25%)	20% (25%)	8% (25%)

consistently make inaccurate predictions about which problems will be most challenging for students.

This unanticipated decoupling between performance on graph interpretation and graph generation is striking. In addition to the large difference in overall accuracy, the relative difficulty of different representations is essentially reversed in the two tasks. Students had the most trouble in interpreting stem-and-leaf plots, but that was the only graph type they had any success in generating. A similar dissociation of interpretation and generation has been observed in other domains such as programming (Anderson, Conrad & Corbett, 1989), although not always (Pennington & Nicolich, 1991).

Histogram & Scatterplot = Bar Graph

When we examined the student graph generation solutions which were correct at least so far as having the proper surface features, we noted a characteristic error in 100% of the histograms and 28% of the scatterplots that provides strong evidence about the students' problem solving strategy. Figure 5 displays a typical histogram solution and Figure 6 displays a typical scatterplot solution. In each case the students have constructed axes that are appropriate for a bar graph. In both graphs, the x-axis represents individual values of a categorical variable (individual brands of peanut butter) and the y-axis represents values of a quantitative variable (sodium level). Each of these graphs is the informational equivalent of a bar graph. This suggests that, in this stage of learning, students are transferring existing knowledge about bar graphs, instead of using knowledge specific to the target representation. That students would already have knowledge of bar graphs is consistent with bar graphs being a simpler representation than histograms (no aggregation of data in the x-axis) and scatterplots (only one continuous variable). This hypothesis not only explains the graph generation results, but also appears to account for the overall pattern of graph interpretation and graph selection results as we discuss below.

Graph Interpretation

Figure 7 displays a set of production rules for common bar graph interpretation problems – given one of the categorical

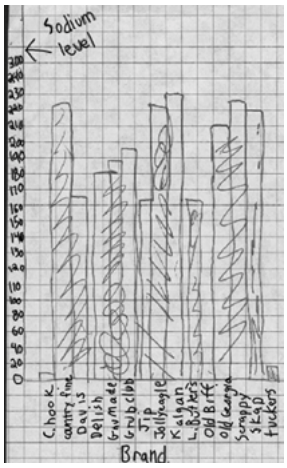


Figure 5: An example of a student-drawn "histogram". Note the axes are a categorical variable versus a quantitative variable – appropriate for a bar graph.

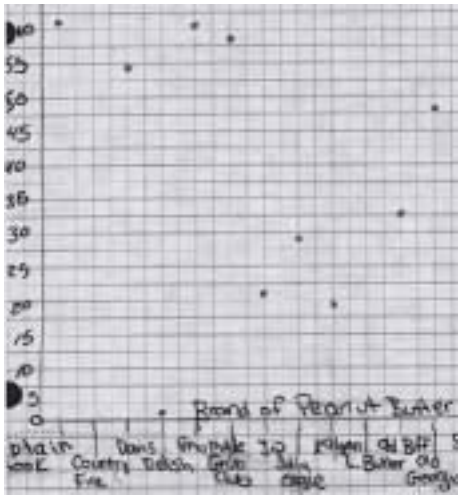


Figure 6: An example of a student-drawn "scatterplot". Note the axes are once again more appropriate for a Bar graph.

instances on the x-axis, find the associated quantitative value on the y-axis. For example, if a bar graph displays the price of several brands of peanut butter (see Figure 4) and a student is asked "What is the price of GnuMade peanut butter?", the student finds "GnuMade" on the x-axis, finds the top of the associated bar, then reads horizontally across to find GnuMade's price on the y-axis. These productions were directly applicable to all of our histogram interpretation exercises. On those exercises, the 15 students had an average accuracy rate of 95%. These productions were also applicable to several of the scatterplot questions (e.g., "What is the price of the brand with a quality rating of 3?"). Students had an average accuracy rate of 50% on these questions.

Student performance on the scatterplot questions is quite striking, when we consider the fact that these questions are not standard for scatterplots. At the same time, the students

Table 2: Average student performance in graph interpretation, for different kinds of problems. Percent which students would be expected to get right through guessing is placed in parentheses where appropriate.

	Histogram	Scatterplot	Stem-and -leaf plot
Interpret: "Bar graph Productions"/ Analogous	95%	50%	21%
Interpretation: Emergent Properties	N/A	67% (50%)	N/A
Interpretation: Overall	95%	57%	17%

had considerable difficulty with other scatterplot exercises. In fact, on the exercises where students had to interpret a scatterplot's global properties, generally considered that representation's main function, the students did not perform significantly better than chance ($p>.10$, $N=12$, using a sign test).

By contrast, the bar graph interpretation productions in Figure 7 do not transfer to stem-and-leaf plots, because the student can neither look up the given value on either axis, nor read the answer off the other axis. This is borne out by the fact that on the 4 questions which were almost word-for-word identical to the histogram interpretation questions, the 13 students performed much more poorly, averaging 21% correct. This is significantly lower than the performance on the corresponding questions for histograms ($t(26)=9.908, p<.0001$).

Hence student interpretation performance levels are more similar between histograms and scatterplots that share surface features with bar graphs, than between histograms and stem-and-leaf plots that have similar content but different surface features (cf. Chi, Feltovich, & Glaser 1981).

Graph Selection

If students are reasoning about all three representations with reference to a single familiar bar graph, they have no basis for discriminating which of the three is appropriate for representing different types of relationships and we would expect chance performance levels. This is in fact what we found, with students getting 15% on the graph selection exercises, even below the 25% accuracy choosing 1 out of 4 would predict. In the absence of understanding which representation is appropriate, an apparent bias in favor of the representation the student had drawn in another question, which by our study design was necessarily wrong, may have led to the observed below chance performance.

```
If we are trying to find a value on a graph
  And we are looking for the value
        corresponding to a value V
Then
  Set a subgoal of looking for V written on the x axis

If we are looking for V on the x axis
  And value V is written at location x* on the x axis
Then
  Set a subgoal of looking at location x*

If we are looking at location x* on the x axis
  And point P is the topmost drawn in graphic
        above location x* on the x axis
Then
  Set a subgoal of looking across from P

If we are looking across from P
  And y* is the value on the y axis horizontally
        over from P
Then
  Return y* as the value we were looking for
```

Figure 7: English-language productions for bar graph interpretation, also suitable for some histogram and scatterplot interpretation.

```
If we are trying to generate a graph G
And G's axes have not been selected
Then
  Set a subgoal for selecting G's axes

If we are attempting to select axes for graph G
  And the X axis is not selected
  And there is a variable V in our data set
  And V is a categorical variable
Then
  Select V as our X axis
```

Figure 8: An English-language production for choosing the X axis variable during generation, overgeneralized to apply inappropriately during scatterplot and histogram generation.

Conclusion and Future Work

Novice students' performance on interpretation, generation, and selection of the data representations in this study can be explained as depending upon transfer of their prior knowledge of bar graphs. Where transfer and generalization are afforded by surface similarity, they occur, whether appropriate or not. This hypothesis exposes a more integrated pattern of interpretation and generation performance than is apparent in the overall results.

Given these findings, we are working toward developing a more complete ACT-R cognitive model of learning data representations

A major subtask in developing our model will be refining our understanding of the factors which scaffold the novices in transferring their knowledge, both appropriately and inappropriately. We do not yet know which common features between representations are essential to this process – certainly it seems that surface features are more important than deeper features, a finding compatible with those in analogical transfer (cf., Novick 1988; Novick and Holyoak 1991) – but which surface features are most salient is an important question in itself. For example, stem-and-leaf plots have three large differences from histograms: flipped axes, the need to remove the tens digit, and the need to count up values. Determining which of them is most

important will have large impacts on our understanding of the generality of the productions students use.

Eventually, we hope to use the cognitive model we develop to build a Cognitive Tutor (Corbett, Koedinger, & Hadley, in press) for this domain. Already, our research has given us extensive information about some of the important difficulties novices have in learning how to generate and interpret these basic representations, including the confusion between histograms and bar graphs. Additionally, our curriculum will be strongly shaped by further research determining whether these overgeneralizations are truly misconceptions which need to be broken down, or whether they are preconceptions which can still be built upon in some way (cf., NRC 1999).

A final future direction is one that may have surprising power – rather than trying to repair misconceptions, we may get even better long-term results from addressing the possibility that we can create a curriculum where bar graphs are taught differently, and the overgeneralization never develops in the first place – where the interpretation productions still transfer, but the generation productions do not become inappropriately broad. Through research in these areas, we hope to transform students' knowledge in this domain.

Acknowledgments

We would like to thank Jay Raspat and Katy Getman for assisting us in the administration of the study discussed here. We would also like to thank Bethany Rittle-Johnson and Jack Zaientz for helpful discussions.

References

Anderson, J.R. (1993) *Rules of the Mind*. Hillsdale, NJ: Erlbaum.

Novick, L.R. (1988) Analogical Transfer, Problem Similarity, and Expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* 14 (3), 510-520.

Scanlon, E. (1998) How Beginning Students Use Graphs of Motion. In M.W. van Someren, P. Reimann, H.P.A. Boshuizen, T. de Jong (Eds.) *Learning With Multiple Representations.* Kidlington, OX UK: Elsevier Science.

Anderson, J.R., Conrad, F. and Corbett, A.T. (1989) Skill acquisition and the LISP Tutor. *Cognitive Science*, 13, 467-505.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121-152.

Corbett, A. T., Koedinger, K. R., & Hadley, W. H. (in press). Cognitive Tutors: From the research classroom to all classrooms. In Goodman, P. S. (Ed.) *Technology Enhanced Learning: Opportunities for Change*. Mahwah, NJ: Lawrence Erlbaum Associates.

Koedinger, K.R. and MacLaren, B.A. (1997) Implicit strategies and errors in an improved model of early algebra problem solving. In Shafto, M.G. & Langley, P. (Eds.) *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society.*382-387.

Larkin, J.H. and Simon, H.A. (1987) Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science,*11, 65-99.

Leinhardt, G, Zaslavsky, O. and Stein, M.K. (1990) Functions, Graphs, and Graphing: Tasks, Learning, and Teaching. *Review of Educational Research.* 60 (1), 1-64.

Nathan, M.J. & Koedinger, K.R. (2000) An investigation of teachers' beliefs of students' algebra development. *Cognition and Instruction.* 18 (2), 207-235.

National Council of Teachers of Mathematics. (2000) *Principles and Standards for School Mathematics.* Reston, VA: National Council of Teachers of Mathematics.

National Research Council. (1999) *How People Learn: Brain, Mind, Experience, and School.* Washington, DC: National Academy Press.

Novick, L.R. and Holyoak, K.J. (1991) Mathematical Problem Solving by Analogy. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* 17 (3), 398-415.

Pennington, N. and Nicolich, R. (1991) Transfer of Training Between Programming Subtasks: Is Knowledge Really Use Specific? *Empirical Studies of Programmers: Fourth Workshop.* 156-176.

Singley, M.K. and Anderson, J.R. (1989) *The Transfer of Cognitive Skill.* Cambridge, MA: Harvard University Press.

Stenning, K., Cox, R., and Oberlander, J. (1995) Contrasting the cognitive effects of graphical and sentential logic teaching: Reasoning, representation, and individual differences. *Language and Cognitive Processes*, 10, 333-354.

Tabachneck, H.J.M., Leonardo, A.M., Simon, H.A. (1994) How Does an Expert Use a Graph? a Model of Visual and Verbal Inferencing in Economics. *Proceedings of the 16th Annual Conference of the Cognitive Science Society.*

Referential Form, Word Duration, and Modeling the Listener in Spoken Dialogue

E. G. Bard (ellen@ling.ed.ac.uk)
HCRC and Department of Theoretical and Applied Linguistics, Adam Ferguson Building, University of Edinburgh
Edinburgh EH8 9LL, U.K

M. P. Aylett (matthewa@cstr.ed.ac.uk)
HCRC and Rhetorical Systems, 2 Buccleuch Place, University of Edinburgh,
Edinburgh EH8 9LW, U.K

Abstract

Referring expressions are thought to be tailored to the needs of the listener, even when those needs might be costly to assess, but tests of this claim seldom manipulate listener's and speaker's knowledge independently. The design of the HCRC Map Task enables us to do so. We examine two 'tailoring' changes in repeated mentions of landmark names: faster articulation and simplified referring expressions. Articulation results replicate Bard et al. (2000), depending only on what the speaker has heard. Change between mentions was no greater when it could be inferred that the listener could see the named item (Expt 1), and no less when the listener explicitly denied ability to do so (Expt 2). Word duration fell for speaker-Given listener-New items (Expt 3). Reduction was unaffected by the repeater's ability to see the mentioned landmark (Expt 4). In contrast, referential form was more sensitive to both listener- (Expt 3) and speaker-knowledge (Expt 4). The results conform most closely to a Dual Process model: fast, automatic, processes let the speaker-knowledge prime word articulation, while costly assessments of listener-knowledge influence only referential form.

Introduction

Speakers are said to design their utterances to suit the needs of their listeners, insofar as those needs can be known (Ariel, 1990; Clark & Marshall, 1981; Gundel, Hedberg, & Zacharski, 1993; Lindblom, 1990). Certainly, there is variation in form. Clarity of pronunciation varies with predictability from local context (Hunnicutt, 1985; Lieberman, 1963) and with repeated mention (Fowler & Housum, 1987). Referential forms are syntactically simpler the more readily interpreted or 'accessible' their antecedents, are (*a blacksmith's cottage* v *it*) (Ariel, 1990, Fowler, Levy, & Brown, 1997; Gundel, et al., 1993; Vonk, Hustinx, & Simmons, 1992). Yet maintaining an incrementally updated model of what the listener knows, what is established common ground, and what the listener needs to know is a considerable cognitive task. Because speaker's and listener's knowledge overlap and because it may be impossible to assess the latter accurately, speakers may default to an account of their own knowledge as a proxy for the listener's (Clark & Marshall, 1981). In fact, many studies simply assume that the two are the same: they manipulate the speaker's knowledge without independently manipulating the listener's (see Keysar, 1997).

This paper compares two versions of the hypothesis that referring expressions are genuinely tailored to the addressee. One deals with the articulation of individual words, the other with the syntactic form of referring expressions. Under current models of language production, NP structure and articulation are generated within units of different sizes, intonational or syntactic phrases on the one hand and phonological words, lexical words, or syllables on the other (Levelt & Wheeldon, 1994; Smith & Wheeldon, 1999; Wheeldon & Lahiri, 1997). Moreover, speech appears to be produced in a cascade, with a sequence of smaller units being prepared for articulation even as the succeeding larger unit is being designed. Thus, incrementally updating a listener model in order to articulate each phonological word appropriately would impose a much heavier computational burden than updating it phrase by phrase. Making both kinds of update for the processes running in parallel would be even more demanding, with the listener model operating both in the state appropriate to the most recently produced word and in the state created by the most recently planned phrase.

We will first develop existing hypotheses about how speakers model listeners while planning and producing speech. Then we will report four studies which test these hypotheses on materials from a single corpus. They follow the comparisons made by Bard et al. (2000) on a psychological measure of clarity, the intelligibility to naïve listeners (recognition rate) of a balanced sample of excised spoken words. The present paper reports a phonetic measure of clarity (word duration), and a syntactic measure of referential form for all suitable cases in a dialogue corpus. Finally, we will discuss the implications of the comparison.

Modeling Listeners while Speaking

Existing accounts of tailoring to listeners' needs make different computational demands on speakers. Where they are not designed with a view to on-line processing, we will attempt to interpret their implications.

Lindblom's *H-and-H Hypothesis* (1990) makes the heaviest computatonal demands. It posits that speakers adjust the articulation of spoken words to the knowledge which the listener can currently recruit to decoding the speech signal: speakers hyper-articulate when listeners lack such auxiliary information and hypo-articulate when redundancy is high. More redundant linguistic environments do contain word tokens articulated with greater speed and less precision (Bard & Anderson, 1983, 1994; Fowler & Housum, 1987; Hunnicutt, 1985; Lieberman, 1963; Samuel & Troicki, 1998). The question is whether this relationship depends on the speaker's consulting an up-to-date model of the listener's current knowledge each time s/he prepares the prosodic character of a phonological word or the articulation of its syllables. Though H-and-H does not preclude defaulting to speaker-knowledge, it is framed in terms of genuine listener-knowledge and implies that speakers should observe listeners continuously for signs of misunderstanding or disagreement. Wherever speaker's and listener's knowledge differ, the latter should take precedence.

In contrast, Brown and Dell (1987) propose a modular division between the initial formulation of utterances and the revision of output which does not adequately convey the intended concepts. The listener's knowledge is implicated only in revision (Dell & Brown, 1991, pp. 119-120). Called the *Monitor and Adjust Hypothesis* (Horton & Keysar, 1996), this model defaults to speaker-knowledge first and pays later – if necessary. As originally formulated, Monitor and Adjust does not explain how the hitherto speaker-driven processes assess the adequacy of an utterance from the listener's point of view. We assume that each interlocutor's knowledge includes a record of what the other has actually said. Listeners' occasional explicit feedback, a minimal listener model, could therefore influence a modular system which revises inadequate utterances. Under this Extended Monitor and Adjust Hypothesis, post-feedback utterances could reflect any listener-knowledge which the feedback has conveyed. Otherwise, listener-knowledge should be irrelevant to production.

The third proposal deals with co-presence, characteristics of listeners which affect likely overlap with speakers' own knowledge (Brennan & Clark, 1996; Fussell & Krauss, 1992; Isaacs & Clark, 1987; Schober, 1993). The manifestations of co-presence in the dialogue literature are many, but the notion was originally used to reduce the computation which a speaker must perform to determine the unknown component of mutual knowledge, i.e. what the listener knows. Under this heading, Clark & Marshall (1981) list shared community membership, physical co-presence of interlocutors and the objects under discussion, and knowledge both of the dialogue and of a suitable scenario. Since much of co-presence is long-lasting, it can reduce both the depth and the frequency of listener modeling. To exploit these economies, speakers should attend to evidence for and against co-presence, and they should maintain defaults for some undefined time after positive evidence. We will call this the *Co-presence Default Hypothesis.*

Finally, Bard et al. (2000) develop a suggestion of Brown and Dell (1987) which we will call the *Dual Process Hypothesis*. It proposes a division between fast, automatic processes, which have no computational cost, and slower, more costly processes requiring inference or attention. The former include priming. (Balota, Boland, & Shields, 1989; Mitchell & Brown, 1988), an effect of the speaker's own recent experience. The latter include the kind of complex reasoning usually implicated in constructing a model of the listener. In competition with listener-modelling are the computations which support planning a dialogue or tracking a shared task. When there is competition for time and attention, (Horton & Keysar, 1996), the inferential processes may suffer, leaving the speaker with only cost-free defaults in the form of his or her own knowledge.

Studies of Intelligibility and Referring Expression

Givenness and Referring Expressions

To test these hypotheses, we made use of two effects of Given status broadly defined. First, spoken words introducing New items are longer and clearer than those in repeated mentions (Fowler & Housum, 1987) but only when the two tokens are co-referential (Fowler, 1988; Bard et al, 1991). Initial mentions of items uttered without visible referents (Prince's (1981) 'brand new') are also longer and clearer than those with visible referents (Prince's (1981) 'situationally' Given) (Bard & Anderson, 1994). Second, referring expressions simplify with repeated mention (*a blacksmith's cottage.... it)* as their antecedents become more accessible (Ariel, 1990, Gundel, Hedberg, & Zacharski, 1993). To compare the two systems, we used a single coded corpus of spontaneous speech which made it possible to select items which were Given to one or both interlocutors on the basis of what each saw, said, or heard in the dialogue.

Method

Materials. Materials came from the HCRC Map Task Corpus (Anderson et al., 1991), 128 unscripted dialogues in which pairs of Glasgow University undergraduates ($N = 64$) communicated routes defined by labeled cartoon landmarks on schematic maps of imaginary locations. Instruction Giver's and Follower's maps for any dialogue matched only in alternate landmarks. Participants knew that their maps might differ but not where or how. Players could not see each other's maps. Familiarity of participants and ability to see the interlocutor's face were counterbalanced. Each participant served as Instruction Giver for the same route to two different Followers and as Instruction Follower for two different routes.

Channel per speaker digital recordings were word-segmented. All words of any expression referring to a landmark were coded for the landmark, tagged for part-of-speech, and parsed. Interrupted or disfluent items were excluded. All remaining expressions making repeated reference to a landmark and meeting experiments' design criteria were used. Duration was measured only if both mentions include the same words. All repeated mentions were are assessed for syntactic form.

Dependent Variables

K-reduction. Normalized duration (Campbell & Isard, 1991) assigns each word token a value, *k*, representing its position in the expected log length distribution for words of its dictionary phoneme composition and stress pattern. *K-reduction* is the difference between the *k*-durations of a read control form and of the corresponding item in running speech. Faster articulation with repeated mention would enhance *k*-reduction.

Form of referring expression. The 27 items with relative clauses in their first mentions were excluded because of a conflict in coding schemes. All other first and second mentions of landmarks (***N*** = 1136) were classed on the scale displayed in Table 1, where '0' indicates least simplified/accessible. Simplification score should increase with repeated mention.

Table 1. Simplification scale for referring expressions

Code	Definition		Example
0	numeral + indef art +	noun sequence	*one mountain* *a mountain*
1	def art + poss +	nominal	*the mountain* *my one*
2	 deictic adj+	possess pro deictic pro nominal	*mine* *that* *this mountain*
3		other pro	*it*

Experiment 1: Inferable Listener Knowledge

Design. Experiment 1 compared repeated mentions of landmarks appearing on both players' maps in two conditions, self- and other-repetition. The key to the design is the fact that a speaker who first mentions a landmark must have it on his or her own map. Thus, in an other-repetition the repeater can easily infer that the introducer can see the landmark. The second token, therefore, refers to an object which is Given both to the repeater who has heard it mentioned and can see it, and to the current listener who has also heard it mentioned, who can see it, and who has mentioned it. Self-repetitions differ in two respects: the repeater who introduced the landmark does not know if the listener can see it. Thus, the design contrasts a case where the listener can easily be concluded to have more knowledge of the referent with one where the listener's knowledge is in doubt. The inference about shared visual resource is both simple and important to the task. Since visibility can affect clarity of mention (Bard & Anderson, 1994), tailoring to the listener here should enhance change across mentions (more *k*-reduction, greater simplification of expression) where the listener has more information - in other-repetition.

Not all hypotheses make this prediction. H-and-H predicts that articulation will be sensitive to the listener's needs in this way. Dual Process predicts instead that any effect will be found in referential form, which is designed over intervals long enough to allow for completing the necessary inference. Copresence predicts effects to what the listener can see. Monitor and Adjust makes no special prediction because speakers are not obliged to model listeners cotinuously to conduct dialogues.

Table 2. Changes with self- v other-repetition:

Measure	Original speaker	
	Self	Other
Articulation: *k*-reduction	0.127	0.192
Form of referring expression	0.878	0.745

Results. Table 2 shows similar changes in articulatory clarity for self- and other-repetition. As in Bard et al. (2000), words were said faster on repeated mention (F_2 (1,691) = 63.75, $p < .0001$) but with no significant difference in reduction between the 263 other-repetitions and the 430 self-repetitions (mention x prior speaker: n.s.). Form of referring expression simplified with repeated mention (F_2 (1,269) = 177.12, $p < .0001$) but again did not distinguish the 90 other-repetitions from the 430 self-repetitions (mention x prior speaker:

n.s.). Contrary to the H-and-H predictions, the listener's experience was not critical. Repetitions of any mentions of visible objects which the repeater had heard were treated alike.

Experiment 2: Listener Feedback

Design. Experiment 2 provides a more direct test of the effects of listener knowledge. When one speaker introduces an unshared landmark, the listener, who lacks it, may provide corrective feedback indicating the discrepancy between the players' maps. Sometimes, however, the listener fails to do this. We compare repeated mentions of the names of unshared landmarks by the same speaker in these two cases. In both, the repeater has said and heard the initial mention and can see the object. When the listener denies having it, the repeater knows that the listener has heard the word but cannot see the object. Otherwise, the repeater cannot tell if s/he can see the landmark.

Cooperative behavior would yield a more restricted effect of repetition where the listener has denied ability to find the object. This comparison is important for the Extended version of Monitor and Adjust, which predicts that feedback at least could make a difference to subsequent mention design. Only Dual Process holds that pronunciation must and syntactic form may be designed without regard to the listener's comments.

Table 3. Changes for repetition with v without feedback on listener's inability to see the landmark:

Measure	Visibility to listener	
	Not denied	Denied
Articulation: *k*-reduction	0.070	0.140
Form of referring expression	0.470	0.410

Results. Table 3 shows that both articulation and form of referring expression were unaffected by feedback. The 73 repetitions with intervening denial and the 122 without abbreviated with repetition significantly and equally ($F_2(1,193) = 9.45$, $p = .0024$; mention x denial: *n.s.*). The simplification of referring expressions on second mention was similar for the 44 cases with intervening denials and the 86 without (F_2 (1,128) = 18.49, $p < .0001$; mention x denial: *n.s.*). Feedback that could block defaulting under Monitor and Adjust does not do so. Only what the repeater has seen, heard, and said seems to play a role.

Experiment 3: Listener Identity

Design. Experiment 3 examines introductory mentions of the same shared landmarks in Givers' two trials with the same map. In the first trial, the landmark is New for both players. In the second, it is Given for the speaker, an Instruction Giver who has mentioned it before, heard that mention, and seen the landmark. However, it is New to each successive listener. Adjustment to the new listener should block any tendency to utter a second introduction as a shorter, more accessible Given item.

In fact, this experiment offers the classic test of Co-presence. Mentions to new listeners should be geared to their ignorance. H-and-H posits that listener modeling will block at least articulatory change. Monitor and Adjust predicts changes in articulation and form, because the speaker's knowledge controls language production, not the listener's. Dual Process predicts a loss of clarity because articulation depends on the speaker's previous mention, not on the listener's knowledge. Only form of referring expression may reflect the listener's ignorance and remain unchanged.

Results. Table 4 shows that second introductions are significantly shorter than first for 239 pairs of words (F_2 (1,238) = 12.48; $p < .0005$). In contrast, simplification of referring expression does not significantly increase over 116 pairs of introductory mentions (F_2 (1,115) < 1). Thus, word reduction appears to reflect the Given status of the item for the speaker, while referential form reflects the fact that the freshly introduced landmark is New for each listener. Greater sensitivity in form of referring expression is predicted only by the Dual Process Hypothesis.

Table 4. Change with reintroductions to new listeners.

Measure	Introduction	
	1	2
Articulation: *k*-reduction	0.498	0.558
Form of referring expression	0.466	0.552

Experiment 4: Speaker Knowledge

Design. In Experiment 4, only other-repetitions were used, but now the landmark in question was either shared by both speakers or absent from the repeater's map. In both cases, the original introducer, who is the listener at the point of second mention, can see the item, has mentioned it, and has heard it mentioned. The repeater has also heard it mentioned, but has not mentioned it and may or may not be able to see it.

Because this experiment holds listener knowledge constant, adjustment to the listener cannot yield any differences between conditions. If the speaker's visual surroundings are important, then changes across

repeated mentions will be greater for visible, shared landmarks than for unshared.

H-and-H predicts no effect of what the speaker can see. Monitor and Adjust allows the speakers' knowledge to affect both dependent variables. Dual Process claims that auditory priming keys articulation to speaker-knowledge, but also allows for costly access to additional information, which would permit effects of speaker-knowledge on referential form.

Table 5. Changes on other-repetition of shared v unshared landmark names

Measure	Visibility to speaker	
	+	–
Articulation: *k*-reduction	0.114	0.183
Form of referring expression	0.745	0.240

Results. Table 5 shows reduction of word tokens with repeated mention (F_2 (1,224) = 12.37, $p < .0005$) but .no significant difference between the outcome for the 144 shared, visible landmarks and the 82 unshared (mention x visibility: *n.s.*). Referential form, however, is speaker-centric. Second mentions are more simplified than first overall (F_2 (1,138) = 24.67, $p < .0001$), but the change is greater for the 90 shared items than for the 50 unshared (mention x visibility: F_2 (1,138) = 6.48, $p < .02$).

This outcome is not consistent with adjustment to listeners alone or with overall use of speakers'-knowledge as a proxy for listeners'. It conforms best to the notion that referential form is sensitive to a wider range of information than articulatory clarity.

Discussion

The experiments reported here and in Bard et al. (2000) test for effects on repeated mentions of several aspects of speaker- or listener-knowledge. Experiment 1 pitted the speaker's own experience in seeing and hearing against the listener's under two conditions, when it could and could not readily be inferred that those listeners could see the landmark. Experiment 2 pitted the speaker's experience of seeing, saying, and hearing against the listener's declared inability to see the item in question. Experiment 3 pitted the speaker's experience in having seen the mentioned landmark, mentioned it, and heard it mentioned against the new listener's ignorance of the item as the landmark was introduced in a second trial. Experiment 4 kept the listener's knowledge constant as well as the speaker's experience in hearing a prior mention, but manipulated the speaker's ability to see the landmark.

In all these cases, the repeating speaker had heard the original mention. In all, clarity of articulation was sensitive only to what the speaker had heard. These are exactly the results found by Bard et al. (2000) for a balanced but restricted sample of materials and with intelligibility to naïve listeners directly measuring clarity. Thus, reduction and consequent reduction in articulatory detail with repeated mention is conditioned by the repeater's experience. There is no indication that models of the listener are consulted.

Referential form showed a different pattern. Like articulation, it was insensitive to some information which should have entered a model of the listener: an indications of what the listener could or could not see (Expts 1 and 2). Yet, it did show two effects which articulation did not. First, referential form did not simplify on re-introduction to new listeners (Expt 3)[1]. In this case, form of referring expression was tailored to the listener's needs. Second, simplification of form across repeated mentions was enhanced when the speaker could see the named landmark (Expt 4). Thus, referential form is more sensitive than articulation but to both interlocutors' knowledge.

Why should form have these characteristics? Form of referring expression does not respond on-line to aspects of co-presence delivered via feedback or inference. It shows neither the complete insensitivity to listeners that Monitor and Adjust predicts for initial design, nor the sensitivity to feedback which should guide redesign. We would argue that Map Task participants juggled competing demands on their attention, as the Dual Process Hypothesis predicts. Unlike the fast automatic processes through which speaker memory affects articulation, slower processes can compete for attention with the communicative task in hand. In this task, however, listener modeling does not take precedence. Only the listener factor most grossly related to the task -- who is participating – affects the design of referring expressions. It shares that honor with an equally basic speaker factor -- what is on the speaker's own map.

The difficulty of the communicative task may well influence the degree to which speakers appear to be modeling their listeners. Certainly direct manipulation of communicative tasks changes speakers' priorities (see Horton & Keysar, 1996). Presumably, speakers could be more sensitive to listener-knowledge if some kind of external record-keeping were to ease the computational burden. The Dual Process Hypothesis predicts that both task and memory load should influence the design of referring expressions, but that neither should affect the articulation of individual words.

[1] While Jurafsky et al. (2001) have recently reported less reduction for reintroduction to new listeners than to old, they find significant reduction to both.

Acknowledgment

This work was supported by ESRC (UK) main grant to the Human Communication Research Centre.

References

Anderson, A. Bader, M., Bard, E.G., Boyle, E., Doherty, G., et al. (1991). The H.C.R.C. Map Task Corpus. *Language and Speech*, *34,* 351-366.

Ariel, M. (1990). *Accessing Noun-Phrase Antecedents*. London: Routledge/Croom Helm.

Balota, D. A., Boland, J. E., & Shields, L. W. (1989). Priming in pronunciation: Beyond pattern-recognition and onset latency. *Journal of Memory and Language*, *28,* 14-36.

Bard, E. G., & Anderson, A. (1983). The unintelligibility of speech to children. *Journal of Child Language*, *10,* 265-292.

Bard, E. G., & Anderson, A. (1994). The unintelligibility of speech to children: Effects of referent availability. *Journal of Child Langua*ge, *21,* 623-648.

Bard, E. G., Anderson, A., Sotillo, C., Aylett, M. Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language*, *42,* 1-22.

Brennan, S., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning Memory and Cognition, 22,* 1482-1493.

Brown, P., & Dell, G. (1987). Adapting production to comprehension -- the explicit mention of instruments. *Cognitive Psychology, 19,* 441-472.

Campbell, W. N., & Isard, S.D. (1991). Segment durations in a syllable frame. *Journal of Phonetics, 19,* 37-47.

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A.K. Joshi, B. Webber, & I. Sag (Eds.), *Elements of Discourse Understanding.* Cambridge: Cambridge University Press.

Dell, G., & Brown, P. (1991). Mechanisms for listener-adaptation in language production: Limiting the role of the "model of the listener." In D. J. Napoli & J. A. Kegl (eds.), *Bridges Between Psychology and Linguistics*. Hillsdale: Erlbaum.

Fowler, C. (1988). Differential shortening of repeated content words produced in various communicative contexts. *Language and Speech, 28,* 47–56.

Fowler, C., & Housum, J. (1987). Talkers' signaling of 'new' and 'old' words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, *26,* 489-504.

Fowler, C., Levy, E. & Brown, J. (1997). Reductions of spoken words in certain discourse contexts. *Journal of Memory and Language*, *37,* 24-40.

Fussell, S., & Krauss, R. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology, 62,* 378-391.

Gundel, J.K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language, 69,* 274-307.

Horton, W., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59,* 91-117.

Hunnicutt, S. (1985). Intelligibility vs. redundancy — conditions of dependency. *Language and Speech*, *28,* 47-56.

Isaacs, E., & Clark, H. H. (1987). References in conversation between experts and novices. *JEP: General, 116,* 26-37.

Jurafsky, D., Bell, A., Gregory, M., Raymond, W., & Girand, C., (2001). Probabilities in the mental grammar: evidence from language production in natural conversation. *Proceedings of CUNY2001*, Philadelphia, PA..

Levelt, W., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition, 50,* 239-269.

Lieberman, P. (1963). Some effects of the semantic and grammatical context on the production and perception of speech. *Language and Speech*, 6*,* 172–175.

Lindblom, B. (1990). Explaining variation: a sketch of the H and H theory. In W. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Keysar, B. (1997). Unconfounding common ground. *Discourse Processes, 24,* 253-270.

Mitchell, D. B., & Brown, A. S. (1988). Persistent repetition priming in picture naming and its dissociation from recognition memory. *JEP:LMC, 14,* 213-222

Prince, E. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), *Radical Pragmatics.* New York: Academic Press.

Samuel, A., & Troicki, M. (1998). Articulation quality is inversely related to redundancy when children or adults have verbal control. *Journal of Memory and Language*, *39,* 175-194.

Schober, M. (1993). Spatial perspective taking n conversation. *Cognition, 47,* 1-24.

Smith, M, & Wheeldon, L (1999). High level processing scope in spoken sentence production. *Cognition*, *73,* 205-246.

Vonk, W., Hustinx, L., & Simmons, W. (1992). The use of referential expressions in structuring discourse. *Language and Cognitive Processes, 7,* 301-33.

Wheeldon, L., & Lahiri, A. (1997). Prosodic units in speech production. *Journal of Memory and Language*, *37,* 356-81

The Utility of Reversed Transfers in Metaphor

John A. Barnden (J.A.Barnden@cs.bham.ac.uk)
School of Computer Science; University of Birmingham
Birmingham, B15 2TT, United Kingdom

Abstract

In metaphor research there is usually some notion of transfer of aspects of the source domain to the target domain. More rarely, transfers in the opposite direction are countenanced, affecting one's perception of source as well as target. This paper argues that, even without this aim, transfers from target to source should happen. One radical claim here is that it is often better to translate information from literal sentences into prevailing metaphorical terms than to translate the information from metaphorical sentences into literal terms. The issues have been obscured by confusion between intuitive directions of static source/target mappings, directions of individual transfer actions, and direction of main intended information flow. Relevance to an implemented AI system for metaphorical reasoning, ATT-Meta, and to Blending Theory are briefly mentioned. Asymmetry of metaphor is also addressed.

Introduction

In this paper, metaphor is thinking or communicating about some target scenario TS in a way that relies on or is motivated by seeing it as something one takes to be qualitatively different from it. Consider: *"In the far reaches of her mind, Anne knew Kyle was having an affair"* (from real discourse: Gross, 1994). TS is what is going on in Anne's mind. Her mind is being seen as a physical space that has "far reaches." The utterance relies on or is motivated by a metaphorical view of MIND AS PHYSICAL SPACE. In this view, the *source domain* is the knowledge domain concerned with physical spaces, locations, etc., and the *target domain* is concerned with minds and mental states/processes. My term *metaphorical view* means much the same as *conceptual metaphor* (Lakoff, 1993). However, I use a different term partly because Lakoff makes claims about conceptual metaphors that do not affect the present paper.

In analyses of metaphor there is usually a notion of *transfer* of aspects of the source domain to the target domain. The transfer involves copying or in some way translating the source aspects. An "aspect" is an entity, property, relationship, proposition, ... that features in source-domain knowledge. The centrality of *source-to-target* (S→T) transfer is especially evident in feature-transfer accounts (e.g., Ortony, 1979) and analogy-based accounts (e.g., Falkenhainer *et al.*, 1989; Holyoak & Thagard, 1989). In the former, understanding of metaphorical utterances is meant to proceed by finding one or more suitable features of a source entity and ascribing them to, or emphasizing them in, a target entity. In analogy-based accounts, the understander either already possesses an S→T mapping handling some aspects of the source or constructs such a mapping on the fly. The understander uses that mapping in transferring *further* aspects of the source to become potentially new target aspects, or at least target aspects that were not previously attended to or not dealt with in the original mapping. The rest of the paper will be geared ostensibly to analogy-based accounts. However, observations to be made generalize to other accounts as long as they involve notions of mapping and/or transfer.

A mapping generally maps more than one aspect of a domain. I call a part of the mapping that is concerned with one specific aspect a *mapping relationship.* Consider the Socrates-as-midwife metaphorical view in Plato's *Theaetetus*, analyzed by Holyoak & Thagard (1989). This involves, amongst others, a mapping relationship between a midwife and Socrates, one between students (even if male) and mothers, one between babies and ideas, and one between pregnancy and idea-development.

Mappings must be distinguished carefully from actions related to them. A mapping is just a set of relationships between aspects of two domains. It is to be distinguished from the act or process of creating the mapping (another possible meaning of the word "mapping"). Also, a transfer is an action. In fact, a transfer generally rests in part on acts of using the existing mapping. For instance, we might transfer the source-domain proposition that *Socrates helps Theaetetus to give birth to [a particular idea]* to become the target-domain proposition that *Socrates helps Theaetetus to produce [that idea].* Here the particular proposition transferred was not previously mapped, but its parts were. There could be more creative transfers as well, such as transferring the cleaning up of the afterbirth to eliminating useless side-effects of a produced idea. Either sort of transfer involves actions exploiting existing mapping relationships (e.g., the one from babies to ideas).

Metaphorical views are generally (and perhaps always) considered to be directed, in a natural and intuitive way. A view of midwives as teachers is clearly distinct from a view of teachers as midwives. This is the asymmetry of metaphor. In a view of A as B, the *direction*

of the metaphor is from B to A. Also, in all mapping-based accounts I have seen, the *intuitive direction of the mapping* is usually taken implicitly or explicitly to be the same as the direction of the metaphor.

But there is a further direction, namely *the direction of information flow*. In most examples in the literature, the information flow is from source to target, in that all that is discussed is informational effects on the target. However, in some accounts, notably interaction accounts (Black, 1979; Waggoner, 1990) and the blending-based account (Turner & Fauconnier, 1995), information flow from target to source (T→S) is allowed for, though much less commonly analyzed than S→T flow. Thus, in general, the direction of the metaphor may not be the same as the direction of all information flow. A major task of this paper is to expand on this point beyond where the literature has taken it so far, showing that it is of much more general importance than heretofore realized.

The plan of rest of the paper is as follows. The next section explains why the direction in which a mapping relationship is used (e.g., during a transfer) is theoretically independent of its intuitive direction. The third section shows argues for T→S as well as S→T information-flow and mapping-usage actions. The fourth section explains that it is inappropriate to think of the direction of mapping always being (wholly) the same as the direction of the metaphor. The fifth section comments on the asymmetry of metaphor. The sixth links the considerations to an implemented AI system, ATT-Meta, that performs some of the reasoning needed in metaphor understanding (Barnden *et al.*, 1996; Barnden & Lee, 1999, 2001).

Directionality: Relationship *versus* Usage

As stated above, mappings are usually described informally as going from source domain to target domain. (Some authors, e.g., Holyoak & Thagard, 1989, occasionally depart from this). Each mapping relationship can just be denoted as an ordered pair *(s,t)* where *s* is the source-domain aspect mapped and *t* is the target-domain aspect it maps to. For instance, in Socrates-as-midwife, a source-domain scenario containing a particular midwife *mw1* is assumed, and we have the mapping relationships *(mw1, Socrates)* and *(give-birth, produce-idea)*.

Now, it is normally assumed that the *use* of a mapping relationship *(s,t)* is in the direction from *s* to *t*. At one point in processing, a source-domain structure involving *s* may be being worked on; and then, typically as a result of an attempted transfer, a structure involving *t* will be considered. But, in principle, the direction of use is independent of the direction of the relationship. If for some reason it were beneficial to create source-domain structures that paralleled existing target-domain ones according to the metaphorical view at hand, a mapping relationship *(s,t)* could be used, in reverse, to go from *t* to *s*. Equally, we could just as well have mapping relationships that intuitively go from target to source without affecting their *usability* from source to target.

In much work on analogy and metaphor, mappings are required (e.g., in SME, Falkenhainer *et al.* 1989) or preferred (e.g., in ACME, Holyoak & Thagard, 1989) to be one-to-one. A one-to-one mapping does not allow there to be two different mapping relationships *(s,t1)* and *(s,t2)* or two different mapping relationships *(s1,t)* and *(s2,t)*. Clearly, if a mapping violated the former condition, some attempted S→T transfers would be faced with extra complication because of the choice between target aspects; and a mapping violating the latter condition would similarly complicate some attempted movements from target to source. These difficulties do not, however, stop a particular *(s,t)* or *(t,s)* relationship in a non-one-to-one mapping being usable in either direction.

Usefulness of Target-to-Source Transfers

Some accounts of metaphor (notably interaction accounts and Blending Theory) allow for T→S transfers. Such transfers therefore use mapping relationships in the T→S direction. However, in those accounts, attention is focused on cases where the ultimate effect is to make some relatively long-term change in the understander's appreciation of the source domain. In contrast, the present paper argues that T→S transfers can be useful even when there is no effect on the source domain that outlives the short-term purposes of the current processing (e.g., understanding a sentence), and where the original goal of the processing is purely to add information to the target domain. The general argument is that T→S transfers can create source-domain information that feeds into within-source-domain processing that in turn ultimately feeds back into some S→T transfer. We therefore have a distinction between *direction of ultimate information flow*, which is normally S→T, from *direction of individual transfers*, which can be T→S, although there must be at least one S→T if ultimate information flow is to be in this direction.

The next three subsections look at different types of T→S transfer.

T→S Transfer: Certainty Downgrading

In a teacher-as-midwife scenario, suppose that reasoning within the source domain establishes, to some level less than absolute certainty, that *Adonis [a student] gave birth to the idea J* (J viewed as a child). Let us call this proposition SP. Suppose that by ordinary S→T transfer this creates the target-domain proposition TP that *Adonis produced J.* But, finally, suppose that there is an independent argument in the target domain that Adonis did not produce J, and that this argument is deemed stronger than the metaphor-based argument. Thus the certainty level ascribed to TP must be downgraded. Now, in discussions of such conflict in the literature, it is not pointed out that *therefore* it may be desirable or perhaps even necessary (a) to downgrade correspondingly the certainty level ascribed to the source-domain proposition (here SP) from which the downgraded target-domain proposition (here TP) came, and possibly also (b) to affirm the negation of SP.

Action (a) is T→S transfer of certainty downgrading. The motivation for it is that the original source-domain proposition (SP) could have been used to support other propositions in the source domain scenario (e.g., that *Socrates has acted as a midwife for Adonis*). The downgrading of SP may therefore be needed so that those other propositions can be properly downgraded—and this may then require withdrawal of earlier S→T transfers of those propositions. Somewhat similarly, affirmation of the negation of SP in the source domain could be useful as it could lead to new inferences in the source domain and hence new S→T transfers.

Metaphorization of the Literal

There is a much stronger and more explicit type of T→S transfer. Consider the following discourse fragment:

> Socrates helped Adonis to give birth to [the idea J]. Similarly, John helped Mary to produce [some idea K].

It would surely be natural to take John's help to be metaphorically a matter of helping a birth. That is, to transfer the target-domain proposition that *John helped Mary produce K* to become the source-domain proposition that *John helped Mary give birth to K*. Then, the rich resources of the source-domain scenario are available to make further inferences. One such inference, albeit an uncertain one, could be that *John acted as a midwife for Mary*. There could then be a further inferences that John was instrumental in introducing Mary to the responsible sexual partner. This matchmaking function of (ancient Greek) midwifery is explicit in *Theaetetus*. Such further propositions could then lead by ordinary S→T transfer to new propositions in the target domain, for instance that John introduced Mary to people who stimulated her ideas.

Notice the contrast to the following method: derive target domain propositions from the first, metaphorical, sentence; then extract target-domain propositions from the second, non-metaphorical, sentence; then integrate the two sets of propositions. Most existing accounts of metaphor, in not properly dealing with the role of metaphorical utterances in *discourse* leave one with the impression that this would have to be the method. But it is highly impoverished compared to the method in the preceding paragraph, as it does not access the rich source-domain scenario. The impoverished method does allow room to infer that John is Socratic, partly because of the word "similarly"; but once the metaphor used in the first sentence has been left behind there is no strong impetus to make that inference, rather than simply interpreting the word to be pointing out that John is like Socrates purely in having helped someone produce an idea.

In any case, the argument does not depend on the word "similarly." The alternative second sentence *"Socrates also helped Theaetetus to produce [idea K]"* could again be felicitously be processed by doing a T→S transfer to get the proposition that *Socrates helped Theaetetus give birth to K*.

The argument somewhat relies on the source domain scenario not having a complete, equally rich and extensive correspondent in the target domain. That is, reasoning in the source domain uses much knowledge about midwives and their role in (ancient Greek) society, where that knowledge is not all mapped to corresponding knowledge in the target domain. That this lack of mapping is common in metaphor is argued further in (Barnden *et al.*, 1996; Barnden & Lee, 2001), but is also linked to common themes in metaphor research such as the relative unparaphrasability of many metaphorical utterances (see, e.g., Waggoner, 1990) and the relative familiarity, richness and accessibility of source domains as opposed to target domains (see, e.g., Lakoff, 1993). Indeed, even if the metaphorical mapping captured all the richness of facts in the source domain, that would still not be enough because methods of reasoning peculiar to the source domain may need to be captured as well. Also, if familiarity of subject matter can affect the facility of people's reasoning even if the pattern of reasoning is kept constant (see, e.g., Johnson-Laird, 1983:29–34), source-domain reasoning stands to have an advantage just from this factor.

Also, it is a mistake to think that if there is an extended use of a metaphor across a stretch of discourse then target-domain information has to be derived from *each* metaphorical patch in the stretch. It may only be necessary to switch to the target-domain once some source-domain *conclusion* has been obtained from within-source-domain reasoning that the stretch stimulates. Thus it may not only be fruitful but also much more economical to metaphorize intervening literal statements than to literalize the metaphorical ones.

In sum, in many cases the proper way to integrate metaphorical discourse elements with non-metaphorical ones is *not* to literalize the metaphorical ones and then do integration (that's the impoverished method above) but rather to metaphorize the literal ones and then integrate. This technique is a radical departure from other research on metaphor, even when discourse-sensitive as in Hobbs (1990). Hobbs does not preclude metaphorization, but he does not appear to have argued for it.

T→S Transfer: Reasoning Queries

One sense in which T→S transfers can occur is through query-directed reasoning (goal-directed reasoning). This is a powerful and important technique in AI generally. In particular, it can be used to focus metaphorical processing fruitfully, and this is especially important given the notorious indeterminacy of the process of extracting information from metaphorical utterances. The ATT-Meta approach places great stress on the technique. In query-directed reasoning, the process starts with a query—a question as to whether something holds. Queries are compared to known propositions and/or used to generate further queries, which, if eventually established, could help provide an answer to the original query; and the in-

vestigation of these queries results in turn from recursive application of the same principle, or by a match to given information. Suppose the discourse at hand is using the Socrates/midwife metaphor, and one reasoning query that has been posted in the target domain (as a result of processing of surrounding discourse) is whether a student *produced a particular idea J.* By virtue of suitable mapping relationships, this query could give rise to the query as to whether the student *gave birth to J.* We can say therefore that the target-domain query has been transferred to become a source-domain query.

Query-directed reasoning of this sort in metaphor or analogy has been advocated by others (e.g., Markman, 1997), and for the particular purposes of the present paper it involves a relatively uninteresting sort of T→S transfer, as it is only useful if it eventually leads to a proposition transfer from source to target. In the student-producing-idea example, the task is to find support for the *student-gives-birth* proposition in the source. This supportedness in the source must be transferred to turn the *student-produces-idea* query in the target into a proposition. Thus, the proposition transfer is S→T even though the query transfer is T→S.

Directionality: Mapping *vs* Metaphor

Given that the intuitive direction of a metaphor of A-as-B is from B to A, it seems obvious that the intuitive direction of the *mapping* involved in the metaphor should be the same. However, this is simplistic—there are exceptions. They may be quite common, and different mapping relationships in a given metaphorical view may intuitively and theoretically be best viewed as going in different directions.

Metaphorical views of IDEAS AS PHYSICAL OBJECTS and MIND AS PHYSICAL SPACE are used in the sentence *"The idea was hidden behind a door in Mary's mind."* Note that whereas the mentioned idea is being viewed as a physical object, so there is indeed a physical object in the source-domain scenario that is mapped to the idea, other implied physical objects in the source-domain scenario, such as the door, should presumably *not* be mapped to ideas. If, therefore, one takes the *property* of being a physical object to map to the property of being an idea, one would have to say that not all applications of *being-physical-object* in the source-domain scenario map to *being-idea* in the target domain. This is an unfortunate gap between the property and its applications, and brings into question the idea that the *being-physical-object* property maps to *being-idea* after all.

One could, of course, say that the *being-physical-object* property does not map at all, and that it is only particular physical objects that map. But that is surely less intuitively appealing than having some sort of mapping of a property. In fact, it is better to take the following, reversed, stance: the property *being-idea* maps to the property *being-physical-object*. This is a natural mapping relationship because it is a reasonable assumption that if one idea in the specific target-domain scenario at hand is being viewed as a physical object then others are too. Consider for instance the following hypothetical discourse segment:

> Many ideas were whizzing around inside Mary's mind. John's question made her think up even more.

This again uses both IDEAS AS PHYSICAL OBJECTS and MIND AS PHYSICAL SPACE. It would be natural to assume that the further ideas implied by the second sentence are also physical objects inside Mary's mind-space (and indeed whizzing around in it). This is another example of metaphorization. Thus, it is plausible that the mapping of *being-idea* to *being-physical-object* has uniform application to all ideas mentioned in the local discourse context.

As for MIND AS PHYSICAL SPACE, the point is yet starker. Consider *"The idea was in the recesses of their minds."* The minds of all the people concerned are being viewed as physical regions. However, *not* all physical regions in the source-domain scenario are being viewed as minds: in particular, the recesses are not. It is therefore much more natural to think in terms of a mind-to-region mapping relationship rather than the other way round.

We can always restrict a not-uniformly-applicable property (or relationship) in such a way that the resulting property (or relationship) does apply uniformly. However, the restriction may be highly unnatural. For instance, in IDEAS AS PHYSICAL OBJECTS the property *being-physical-object*, restricted to apply only to a physical object *that happens to correspond to a particular mentioned idea* trivially coheres with its applications, with respect to to the metaphorical mapping. Of course, the italicized restriction is not itself a restriction framed in terms of the source domain.

Given that a metaphorical mapping typically involves several mapping relationships (e.g., applying to different properties), the possibility arises of having its component relationships go in different directions. Also, a given mapping relationship together with its inverse may both be intuitively part of the mapping. To accommodate these possibilities we can either broaden the notion of mapping to allow relationships to go in different directions, or we can replace the single mapping by two mappings, one consisting of S→T relationships and the other of T→S ones.

Asymmetry of Metaphor

It is frequently pointed out (e.g., Way, 1991) that metaphor is asymmetric (see Introduction above). The present paper might be thought to conflict with this, as it claims that the direction of the metaphor does not completely determine the direction of transfers or even the intuitive direction of mapping relationships. However, there is no conflict, because of the following points.

There must still be some information flow from source to target, i.e. in the direction of the metaphor: it is the *target* domain that contains the topic being attended to. Even if teachers-as-midwives and midwives-as-teachers

both happened to use exactly the same mapping relationships (up to inversion), it is very different to conclude, say, that in reality certain teachers help certain students from concluding, say, that in reality certain midwives help certain pregnant mothers.

But, in any case, the present paper in no way implies that A-as-B would indeed involve exactly the same mapping relationships as B-as-A, even though in practice there may well be considerable overlap, especially as a structural isomorphism between parts of two domains is an inherently symmetric thing. The isomorphism that is appropriate for one direction of metaphor may be slightly or greatly different from the one appropriate to the other direction. (For one thing, even with a fixed direction there can be competing possibilities for partial isomorphism. Such competition is an important aspect of SME and ACME.) Whereas in some particular discourse a use of teachers-as-midwives might involve an isomorphism between the process of giving birth and the process of producing an idea, a use of midwives-as-teachers in another discourse might rest on an isomorphism between the process of a mother coming to bond with her already-born baby and the process of a person producing an idea.

Finally, mapping relationships for A-as-B and B-as-A could be intuitively similar but be different in detail. We saw that for IDEAS AS PHYSICAL OBJECTS, not all physical objects in the source-domain scenario are mapped to ideas, whereas all ideas in the target-domain scenario are likely to be mapped to physical objects. If a metaphorical view of PHYSICAL OBJECTS AS IDEAS were to be used in discourse, it could similarly be that not all ideas in the source-domain scenario were mapped to physical objects but that all physical-objects in the target-domain scenario were mapped to ideas. Asymmetry is addressed again at the end of the next section.

The ATT-Meta System

The ATT-Meta system is too complex to be described at any length here. It is described in Barnden & Lee (1999, 2001). The present section summarizes how the system is related to the issues in previous sections.

The system is aimed at performing the reasoning needed for understanding a broad class of metaphorical utterances that we call *map-transcending utterances based on familiar metaphors.* Here the understander already knows the metaphorical views used, and therefore possesses source/target mappings underlying those views; however, the utterance uses aspects of the source domains(s) that are not mapped by the known mappings. The system is designed on the principle that there should by default be no attempt to create new mapping relationships to handle those aspects; rather, the system should try to do reasoning that links those aspects to source aspects that are already mapped. For example, consider again the sentence *" In the far reaches of her mind, Anne knew Kyle was having an affair."* ATT-Meta handles this as follows, given knowledge of MIND AS PHYSICAL SPACE and IDEAS AS PHYSICAL OBJECTS, and most importantly the knowledge that (in)ability to operate physically on an idea, in the source domain of IDEAS AS PHYSICAL OBJECTS, maps to (in)ability to operate mentally on the idea, in the target domain. Assume that ATT-Meta's mappings do not map the far-reaches location within a space to anything, so that the utterance transcends the system's mappings in this respect. ATT-Meta can reason, using (mainly) commonsense knowledge about physical spaces and objects, that Anne has only a very low ability to operate physically on the idea that Kyle was having an affair. (This is because the far reaches of a physical region are very distant from the main part of the region, and Anne, or rather her conscious self, is taken to be in that main part.) Then, using the known mapping, ATT-Meta can infer that Anne has only a very low ability to operate mentally on the idea. This example is treated in much more detail in Barnden & Lee (2001). A variety of other examples are also treated in that report and in other reports cited in it.

ATT-Meta's long-term domain knowledge and its knowledge of metaphorical views is couched in terms of IF-THEN rules. A given metaphorical mapping relationship takes the form of a rule, such as (roughly)

> *IF J is in reality an idea AND J is being viewed as a physical object AND person X is being viewed as being able to operate physically on J to at least degree D*
> *THEN in reality X can operate mentally on J to degree at least D.*

ATT-Meta's reasoning is entirely query-directed. So, for instance, in the Anne/Kyle example the reasoning steps mentioned arise from a backwards-going process of query construction, proceeding backwards through rules. In particular, a query about the degree of ability of Anne to operate mentally on the Kyle-affair idea in reality leads to the creation also of a query about the degree of ability of Anne to operate physically on the idea, under the metaphorical view. Thus, the system exhibits T-to-S transfer of queries.

The system's metaphor-based reasoning is thoroughly integrated into a general-purpose framework for qualitatively uncertain reasoning. Reasoning in source-domain terms and in target-domain terms is generally uncertain. Rules and propositions are annotated with qualitative certainty levels, and there is a heuristic conflict-resolution mechanism that attempts to adjudicate between conflicting arguments. As a result of conflict-resolution, the certainty of one or more propositions is downgraded. Reasoning leaves behind a record of dependency links between propositions, so certainty downgrade of a proposition leads to downgrading also of propositions dependent on it. Now, for a given S→T mapping relationship there is often a converse mapping relationship, e.g. (to continue the above example),

> *IF J is in reality an idea AND J is being viewed as a physical object AND in reality person X can operate mentally on J to at least degree D*

THEN X is being viewed as being able to operate physically on J to degree at least D.

Consequently, ATT- Meta performs T→S transfer of certainty downgrades when suitable target-domain propositions are downgraded. Because of the extensive within-source reasoning that ATT-Meta often performs, downgrade within the source domain can lead to other downgrades there by propagation along dependency links.

Because of the T→S mapping relationships, ATT-Meta can metaphorize literally-stated information, and such metaphorization steps are seamlessly mixed in with other reasoning steps. However, only limited experimentation on this has been done so far using ATT-Meta.

As for asymmetry of metaphor, it is instructive to look at the following situation that could hold in ATT-Meta. Recall the T→S rule mentioned above for IDEAS AS PHYSICAL OBJECTS. It can be reworded to have the overall form:

IF J is in reality an idea AND J is being viewed as a physical object AND ... THEN ... .

Now consider S→T rules for a view of PHYSICAL OBJECTS AS IDEAS. These will have the overall form:

IF O is in reality a physical object AND O is being viewed as an idea AND ... THEN ... .

The rule-forms differ crucially in their first two conditions. Thus, T→S rules for A-as-B will be act very differently from S→T rules for B-as-A, even when the source and target aspects mapped are the same or similar.

Conclusions

The main conclusions are that (a) target-to-source transfers of several distinctly different types are desirable in metaphorical discourse understanding; (b) in particular, metaphorization of within-target-domain utterances can be desirable; and (c) the directions of mapping relationships are sometimes intuitively the wrong way round in accounts of metaphor.

The ATT-Meta system is one that routinely allows target-to-source transfers of the sorts mentioned above. Blending theory (cited earlier) allows transfers into and out of the blended space, including transfers from the blended space back into a source domain. Also, the LISA model for analogy (Hummel & Holyoak, 1997) allows mapping to go from target to source, though the relationship to the considerations in the present paper is unclear. Thus, a small number of approaches are beginning to allow effects such as those in this paper. However, the topic appears to have seen little psychological experimentation or computational realization, and would be a fertile ground for future empirical investigation.

Acknowledgments

This research is supported by grant GR/M64208, Engineering and Physical Sciences Research Council.

References

Barnden, J.A., Helmreich, S., Iverson, E., & Stein, G.C. (1996). Artificial intelligence and metaphors of mind: within-vehicle reasoning and its benefits. *Metaphor and Symbolic Activity, 11* (2), 101–123.

Barnden, J.A., & Lee, M.G. (1999). An implemented context system that combines belief reasoning, metaphor-based reasoning and uncertainty handling. In P. Bouquet, P. Brezillon & L. Serafini (Eds), *Lecture Notes in Artificial Intelligence, 1688*.Springer.

Barnden, J.A., & Lee, M.G. (2001). Understanding usages of conceptual metaphors: An approach and artificial intelligence system. Tech. Rep. CSRP-01-05, School of Computer Sci., Univ. of Birmingham, UK.

Black, M. (1979). More about metaphor. In A. Ortony (Ed.), *Metaphor and Thought*. Cambridge, UK: Cambridge Univ. Press.

Falkenhainer, B., Forbus, K.D., & Gentner, D. (1989). The Structure-Mapping Engine: algorithm and examples. *Artificial Intelligence, 41* (1), 1–63.

Gross, L. (1994). Facing up to the dreadful dangers of denial. *Cosmopolitan, 216*(3), USA ed.

Hobbs, J.R. (1990). *Literature and cognition.* CSLI Lecture Notes, No. 21, Stanford University.

Holyoak, K.J. & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science, 13* (3), 295–355.

Hummel, J., & Holyoak, K. (1997). Distributed representation of structure: A theory of analogical access and mapping. *Psychological Review, 104* (3), 427–466.

Johnson-Laird, P.N. (1983). *Mental models: towards a cognitive science of language, inference and consciousness.* Cambridge, MA: Harvard University Press.

Lakoff, G. (1993). The contemporary theory of metaphor. In A. Ortony (Ed.), *Metaphor and Thought*, 2nd ed. Cambridge, UK: Cambridge Univ. Press.

Markman, A.B. (1997). Constraints on analogical inference. *Cognitive Science, 21* (4), 373–418.

Ortony, A. (1979). The role of similarity in similes and metaphors. In A. Ortony (Ed.), *Metaphor and Thought.* Cambridge, UK: Cambridge Univ. Press.

Turner, M., & Fauconnier, G. (1995). Conceptual integration and formal expression. *Metaphor and Symbolic Activity, 10* (3), 183–204.

Waggoner, J.E. (1990). Interaction theories of metaphor: psychological perspectives. *Metaphor and Symbolic Activity, 5* (2), 91–108.

Way, E.C. (1991). *Knowledge representation and metaphor.* Dordrecht: Kluwer.

A Model Theory of Deontic Reasoning About Social Norms

Sieghard Beller (beller@psychologie.uni-freiburg.de)
Department of Psychology, University of Freiburg
D-79085 Freiburg, Germany

Abstract

This paper outlines a model theory of deontic reasoning. It proposes that social norms form the basic concept on which deontic inferences operate. The theory unifies and extends current deontic approaches. Empirical findings from the deontic selection task will be presented which support the theory.

Introduction

Deontic reasoning is thinking about which action a person *may* or *must* perform with respect to a social rule. Imagine an officer who has to administer the admission to an event. The promoter has given him the rule "If a person has a ticket, then this person *may* enter." Lisa has *no* ticket. *May* she enter? The officer answers: "No, Lisa *must not* enter. She does not fulfill the entry condition." Although intuitively plausible, this answer is in conflict with standard conditional logic. The antecedent of a conditional '*If* P *then* Q' denotes a sufficient but not necessary condition. If 'P' does not hold then one can *not* decide whether 'Q' holds. Thus, the officer should rather ask his promoter what to do with Lisa instead of refusing the entry. Why is it, that his answer nevertheless seems right? The goal of this paper is to introduce a new theory of deontic reasoning which explains such phenomena. It will be compared to alternative approaches and backed up with findings from the deontic "micro-laboratory" of the selection task (Wason, 1966).

The Deontic Mental Models Theory (DMM)

The model theory assumes that reasoning requires the construction of mental models which represent the meaning of, for example, verbal premises or a person's background knowledge (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991). Their structure and content capture semantic relations of the situations they refer to. To make an inference, reasoners first extract a putative conclusion from an initial model and then validate it or, if necessary, revise it by fleshing out alternative models. Of which structure and content are the models that underlie deontic reasoning? In this article, it is proposed that the underlying models represent social norms.

Normative Models and Deontic Inferences

Social norms constitute constraints on actions but they do not really restrict a person's freedom of action. A person may follow a norm or violate it. Consequently, two types of models must be distinguished: factual models and normative models.

Factual models describe which conditions hold or which actions are taken. A condition (`C`) is viewed as a state of affairs that can be fulfilled or not (symbolized as `C` vs. ¬`C`). Actions are taken by a person on a certain occasion; not performing an action is notated by negation (`Action` vs. ¬`Action`); a more detailed analysis is given by von Wright (1963). The fact, for example, that a person with a ticket is entering an event can be represented by the following factual model:

(1) Normative	Factual
	[entering] [ticket]

Normative models cannot describe which actions persons really take. They describe constraints on actions, that is, under which conditions actions are forbidden or obligatory. In a consistent system of norms an action cannot be forbidden and obligatory at the same time, otherwise the person is trapped in a dilemma. In the following, 'bans' are taken as the basic normative concept and represented as `forbidden(Action)`. In accordance with the axiom of definitional equivalence in modal logics '*must* X ≡ *must-not* ¬X' (e.g., Chellas, 1980), an obligation can also be represented as a ban: the obligation to take an action means that the omission of the action is forbidden (`forbidden(`¬`Action)`).

With regard to the relation between a ban and its conditions, two assumptions are made: First, people represent *each* ban together with *all* conditions that put the ban into force (*closed world assumption*). Second, people treat the relation between a banned action and the conditions as an *equivalence* (which is justified under the closed world assumption): If the conditions are met, then the action is forbidden, otherwise it is allowed. Taking both assumptions together, the basic schema of a norm (concerning one action) takes the following explicit standard form:

(2) Normative		Factual
[forbidden(Action)]	[Conditions]	
[¬forbidden(Action)]	[¬Conditions]	

Each line denotes a separate model. Since all conditions concerning this action are considered (closed world assumption), the representation is exhaustive (indicated by square brackets). In the simplest case, there is one norm with one condition. Having a ticket, e.g., is often the only condition to be admitted to an event:

(3) Normative		Factual
[forbidden(entering)]	[¬ticket]	
[¬forbidden(entering)]	[ticket]	

Sometimes, however, several conditions have to be considered in combination. Necessary conditions may be treated by a conjunctive model, alternative conditions by a disjunction of models. Spectators of German soccer matches, for example, are often examined not only for their tickets but additionally for weapons. The normative models then contain two disjunctive conditions: A person is not allowed to enter, if (and only if) he or she has no ticket *or* has a weapon:

(4) Normative		
[forbidden(entering)]	[¬ticket]	[weapon]
[forbidden(entering)]	[¬ticket]	[¬weapon]
[forbidden(entering)]	[ticket]	[weapon]
[¬forbidden(entering)]	[ticket]	[¬weapon]

Of course, there may be additional norms concerning other actions as well. Their representation, however, follows the same schema. *Deontic inferences* connect normative and factual models. If a person fulfills the conditions (Cs) associated with a forbidden action then one can assert "the action *must not* be taken" and – according to the axiom of definitional equivalence – "the action *must* be omitted". Thus, two inferences can be drawn from the corresponding set of models[1]:

(5) Normative		Factual
[forbidden(Action)]	Cs	[Cs]
...		∴ must-not Action
		∴ must ¬Action

If an action is taken which is potentially forbidden, then it follows that the conditions Cs *must not* be fulfilled or else the norm would be violated. Equivalently, it *must* be the case that the conditions are *not* fulfilled:

(6) Normative		Factual
[forbidden(Action)]	[Cs]	[Action]
		∴ must-not Cs
		∴ must ¬Cs

While the modals *must not* and *must* correspond to the notions of *ban* and *obligation*, the modals *may* and *need not* are related to the concepts of *permission* and *release* from obligation. Both pairs, ban and permission as well as obligation and release, are *contradictories*. In a consistent system of norms, exactly one of each pair is true: an action is either forbidden or allowed; it is either obligatory or not. One can infer that something *may* be the case if it is not forbidden, and that something *need not* be the case if it is not obligatory. Finally, a norm is *violated* if a person takes an action while fulfilling at least one condition under which the action is banned:

(7) Normative		Factual	
[forbidden(Action)]	[Cs]	[Action]	[Cs]
		∴ violation	

To illustrate the application of the theory, let's reconsider the introductory example. An officer was given the rule: If a person has a ticket, then this person *may* enter. It mentions one condition for the action of entering. Since norms constitute constraints on actions, the officer can map the rule to norm (3) which expresses that entering is not forbidden with a ticket, but it is forbidden without one. For Lisa who has no ticket, model (5) applies which allows the officer to answer: "She *must not* enter".

"Why a New Theory of Deontic Reasoning?"

... one may ask since a number of well-established proposals already exist. While each of the current approaches emphasizes a different aspect, the proposed DMM theory tries to unify their main characteristics.

DMM theory takes up two previous ideas: (1) Modal terms gain their deontic meaning by referring to deontic norms (Johnson-Laird, 1978) which represent (2) permissible and impermissible situations (Johnson-Laird & Byrne, 1992). DMM theory goes beyond these ideas by proposing a concrete representation of norms and relating deontic modals to it. Manktelow and Over (1991; 1995) claimed that social roles and utilities need to be incorporated. These factors are indeed important since social roles distinguish between the parties affected by a social rule and utilities influence its negotiation. They seem not necessary, however, for deontic inferences (Johnson-Laird & Byrne, 1992). Once a social rule has been established, it defines the normative constraints on each parties' actions, and the corresponding normative models determine the possible deontic inferences.

Thompson (2000) argued that a theory of reasoning should not only specify the inferential procedures that operate on a given representation; it must also specify the interpretative processes that set up this representation. As her experiments show, the interpretation of conditional reasoning tasks is affected by two factors: by the perceived necessity and sufficiency relations and by the pragmatic relation (deontic vs. factual). DMM theory integrates necessity and sufficiency relations on the condition-side of norms and it considers the characteristics of the deontic domain she condensed from her studies: the normative character of the models which gives relevance (Sperber, Cara & Girotto, 1995) to the notion of norm-violation. In addition, DMM theory may be used to analyze the interpretation of normative statements by exploring how they are related to norms.

The theory of *pragmatic reasoning schemas* (PRS;

[1] For reasons of simplicity, the action side of the norm is represented exhaustively: all banning conditions are subsumed under Cs. The condition side in model (5) is not represented exhaustively because there may be other bans under the same conditions.

Cheng & Holyoak, 1985; Holyoak & Cheng, 1995) proposed two deontic inference schemas – one for permission and one for obligation – each consisting of four rules which are applicable when their appropriate content is present. The rules of the permission schema are:

P1: If the action is to be taken, then the precondition *must* be satisfied.
P2: If the action is not to be taken, then the precondition *need not* be satisfied.
P3: If the precondition is satisfied, then the action *may* be taken.
P4: If the precondition is not satisfied, then the action *must not* be taken.

The two deontic schemas are sufficient to explain many findings with deontic reasoning tasks (see, e.g., Holyoak & Cheng, 1995). As a theory of deontic reasoning, however, the PRS approach is faced with two problems. First, with regard to terminology the two schemas are not clearly distinguishable – both include a permission rule (e.g., P3) and an obligation rule (e.g., P1) – and the modal terms defining the schemas are themselves undefined (Manktelow & Over, 1995). The idea of a domain-specific representation of norms is adopted by DMM theory but it uses a single normative concept (bans) instead of different schemas and defines the modals by reference to norms, actions, and conditions. This approach encompasses both PRS schemas (Beller, 1997). Second, the scope of PRS theory is quite restricted: It does not cover some deontic inferences that people easily draw. For example, from the entry-rule "If a man has no ticket, then he *must not* enter" people easily infer that without a ticket "he *must* stay out". A corresponding inference rule is available in neither schema; the inference is supported, however, by DMM theory (model 5 applied to norm 3). By considering relations of modal logics the range of covered deontic inferences is extended beyond the PRS schemas.

"What is the *origin* of domain-specificity in reasoning?" is the question posed by evolutionary approaches. Are domain-specific concepts learned as assumed, for example, by PRS theory or do they reflect innate evolutionary adaptations (e.g., Cosmides, 1989; Cummins, 1996)? DMM theory stresses which information persons represent in their models and how these affect reasoning. It is open with respect to the origin question.

Having justified the theoretical relevance of the new theory, it is now applied to Wason's (1966) selection task. Since the discovery of content effects in the 1970s, this prominent paradigm has developed into a micro-laboratory of deontic reasoning with findings that each deontic theory must be able to explain.

Touchstone "Deontic Selection Task"

How does DMM theory fit the basic findings with deontic task versions? For reviews of the vast selection task literature see, for example, Beller (1997), Evans, Newstead and Byrne (1993), or Newstead and Evans (1995).

The Deontic Task: In the original, non-deontic task (Wason, 1966) persons are shown four cards with a letter on one side and a number on the other side. One side is visible: "A", "K", "5", and "8". A rule is given: "If there is an 'A' on one side, then there is a '5' on the other." The task is to select all cards which need to be turned over to find out whether the rule is true or false. Since a conditional '*If* P, *then* Q' is false only if the antecedent 'P' holds and the consequence 'Q' is false, exactly two cards can prove the rule: the 'P'-card ("A") and the '*not* Q'-card ("8"). This answer is usually given by less than 10 % of the participants (e.g., Evans, Newstead & Byrne, 1993). Now, consider the following deontic version (adapted from Griggs & Cox, 1982):

> *Imagine that you are a police officer on duty. It is your job to ensure that people conform with certain rules. The cards in front of you have information about four people. On one side of a card is a person's age and on the other side is what the person is drinking. Here is a rule: If a person is drinking beer, then he or she must be over 19. Select the card(s) that you need to turn over to determine whether people are violating the rule.*

The cards show: "drinking beer", "drinking Coke", "22 years", and "16 years" ('P', '*not* P', 'Q', and '*not* Q' with respect to the rule '*If* P, *then must* Q'). As in the abstract task, 'P'/'*not* Q' should be selected because these cards indicate a rule violation: a person under 19 drinking beer. Deontic tasks often yield solution rates of about 70-90 % (Dominowski, 1995). Different from the abstract task, people need not to evaluate the truth of the conditional. According to DMM theory they can construct normative models that tell them which persons (cards) they have to check. With the closed world assumption and the equivalence assumption, the drinking age rule can be mapped to norm (8):

```
(8) Normative
 [forbidden(drinking_beer)]      [¬over_19]
[¬forbidden(drinking_beer)]       [over_19]
```

The norm is violated (model 7 applied to 8) by a person under 19 ('*not* Q') who is drinking beer ('P') which can be checked for by selecting the cards 'P'/'*not* Q'.

It was this "facilitation effect" that necessitated a deontic theory. Subsequent experiments revealed several factors that are of particular relevance for the deontic solution. Besides the use of the deontic term *must* (e.g., Platt & Griggs, 1993) and a "detective" scenario (van Duyne, 1974) – both strengthening the deontic interpretation – three factors received particular attention: the instruction, the type of negation used, and the rule form.

Instruction: While the abstract version asks for testing the truth of the conditional, the deontic task requires to detect cases of rule violation thereby making clear that each card has to be examined independently from the others (Stenning & van Lambalgen, in press). The high rate of 'P'/'*not* Q' in deontic tasks decreases when the testing instruction is used (e.g., Noveck & O'Brien, 1996; Yachanin, 1986). This is exactly what one would expect from the perspective of a deontic theory, because the testing instruction is not applicable in the deontic

case. Different from indicative conditionals, the truth of a deontic rule is independent from individuals who may conform to the rule or not. Its truth cannot be determined by simply observing persons' behavior – little astonishing that the solution rate decreases.

Negation: The use of explicit negation turned out to be crucial for the solution of tasks with abstract deontic rules like "If one is to take action 'A', then one must first satisfy precondition 'P' " (e.g., Cheng & Holyoak, 1985). An explicit negation of the fact that a person "has taken action 'A' " would be "has not taken action 'A' " while the statement "has taken action 'B' " can be regarded as an implicit negation. Typically, the proportion of 'P'/'*not* Q' decreases when implicit negation is used on the cards (e.g., Jackson & Griggs, 1990; Noveck & O'Brien, 1996). But again, this is consistent with a deontic theory as Holyoak and Cheng (1995) pointed out. Two actions need not exclude each other; they can take place at the same time. If a reasoner does not know whether taking action 'B' and action 'A' are mutually exclusive there is no basis to interpret the two "implicitly negated" cards '*not* P' and '*not* Q' as really negated.

Rule Form: Persons' apparent insensitivity to syntactic modifications of the conditional rule used in the tasks has been taken as an argument against a purely "syntactic" view of reasoning. Cosmides (1989), for example, reversed conditionals from '*If* P *then must* Q' in the 'standard' form to '*If* Q *then* (*may*) P'. From a syntactic point of view, one may expect that the cards to be selected should switch correspondingly from 'P'/'*not* Q' to '*not* P'/'Q'. From a deontic point of view, the reversed rule cannot be violated at all (in the sense of doing something forbidden) because the consequence (by using the modal *may*) does not express a behavioral constraint a person could offend. Consequently, none of the cards should be selected. In either case the predominant selection should change. Empirically, the opposite has been found: 60-70 % keep choosing 'P'/'*not* Q' (e.g., Cosmides, 1989). What is the reason for that? Consider two drinking-age rules: The standard form "If a person is drinking beer (P), then he or she *must* be over 19 (Q)" and the reversed one "If a person is over 19 (Q), then he or she *may* drink beer (P)". According to DMM theory, both rules can be mapped to the same norm (8) although they describe different aspects. The norm is violated by a person drinking beer who is not over 19 ('P *and not* Q'). By assuming that people derive their answer from this norm and not from the conditionals, DMM theory accounts for the insensitivity to their form. The related effect of perspective change (e.g., Gigerenzer & Hug, 1992; Manktelow & Over, 1991) can be explained in a similar way (Beller & Spada, 2000).

Deontic Tasks Without Deontic Solution? The previous results can all be brought in line with DMM theory. Cosmides (1989), however, reported findings that seem to rule out a deontic explanation categorically. She demonstrated that non social contract (non-SC) versions of deontic standard rules produce significantly less facilitation than equivalent social contract (SC) versions although both are said to trigger the same deontic solution. One of her examples is the 'school problem' about assigning students to either Grover High or Hanover High. Both versions mention the deontic conditional: "If a student is to be assigned to Grover High School, then that student *must* live in Grover City". However, while the SC problem (task 9) specifies that the cards should be checked for *cheating*, the non-SC version (task 10) leaves the subjects with an incomplete deontic interpretation. In this latter task it is said that "There are *rules* to determine which school a student is to be assigned to, the most important of these rules is …". Rule violations are attributed to an "… old lady … who often made mistakes when categorizing student documents" (p. 270). The first quotation implies that several normative rules have to be applied in the categorization process but not all of them are known (the closed world assumption is violated). The term "mistakes" leaves open whether the old lady incorrectly assigned students to Grover High *or* Hanover High. – Experimental manipulations that weaken the deontic interpretation or end up with an inconsistent or incomplete interpretation may result in a decrease of the deontic solution but they cannot be taken as an argument against a deontic explanation.

Rule-Change Revisited

The DMM explanation of the deontic selection task assumes that persons do not rely on the conditional rule itself but on normative models that tell them which cards to check. The finding with switched rules corroborates this hypothesis. Nevertheless, persons could also have derived their solution from the conditional rule since there is a rule available in both the standard and the switched version. A stronger argument in favour of the "normative-models hypothesis" would be, if (1) people kept choosing the same cards in a "rule-free" selection task – like those that have been used recently to back up the effect of knowledge about causal relationships (Beller & Spada, 2000) and about promises (Beller & Spada, 2000; Fiddick, Cosmides & Tooby, 2000). The rule-change effect seems to show subjects' insensitivity to the form of deontic conditionals. Although both the standard and the switched form are consistent with one and the same norm, the switched rule cannot be violated deontically as argued above. Thus, if (2) a task does not allow persons to construct a normative model but requires to evaluate the conditional itself, then they should state correctly which conditional can be violated. Both hypotheses are examined in the following experiment (Beller, 1997).

Materials: The materials comprised six tasks: five versions of selection tasks and one rule evaluation task. Four deontic conditionals were used:

R1: If a child is drinking Coke,
then he or she *must* be over 12 years of age.

R2: If a child is over 12 years of age,
then he or she *may* drink Coke.

R3: If a child is over 12 years of age,
then he or she *need not* drink juice.

R4: If a child is drinking Coke,
then he or she *must not* be under 12 years of age.

R1 denotes a standard rule ('*If* P *then must* Q' with "drinking Coke" symbolized as 'P' and "over 12 years" as 'Q') while R2 is reversed ('*If* Q *then may* P'). R3 and R4 are corresponding rules using the other two deontic operators. All rules are derived from norm (9) that forbids drinking Coke if a child is younger than 12 years:

(9) Normative	
[forbidden(drinking_coke)]	[¬over_12]
[¬forbidden(drinking_coke)]	[over_12]

The *selection tasks* started with an introductory part:

> *In a particular country there are two beverages popular with children: Coke containing caffeine and a particular sort of juice. A scientific study has shown that the circulatory system of children younger than 12 years is often affected by drinking Coke. Therefore, the government passed a rule permitting to drink Coke dependent on age. A dispenser offers both beverages, the juice and Coke. The children of a school class are standing in front of the dispenser together with their teacher. Some of the children are already over 12 years of age, some are under 12 years. The teacher reminds her pupils of the correct behavior. [She mentions the following beverage rule].*

The five versions differed in the rule following right after this part: Four tasks mentioned one of R1-R4; the fifth task comprised no conditional (and omitted the sentence in square brackets). The instruction continued:

> *The cards below represent four children who took a beverage from the dispenser. On one side of each card it is written whether the child is drinking Coke or juice, the other side shows whether he or she is over 12 years. Your task: Please indicate all the cards that you would have to turn over (i.e., all of which you need to know the information on the back) in order to find out whether the child has violated the beverage rule.*

The cards read: "is drinking Coke", "is drinking juice", "is over 12", and "is under 12" ('P', '*not* P', 'Q', and '*not* Q'). Since all tasks can be mapped to one and the same norm (9), the model theory predicts that people choose the same cards 'P'/'*not* Q' that may indicate a violation of this norm (model 7 applied to 9) by children under 12 ('*not* Q') who are drinking Coke ('P').

The instruction of the *rule evaluation task* required to evaluate the deontic conditionals (R1-R4) without reference to an underlying norm:

> *Please read carefully through the following if-then-statements. Check for each statement whether it expresses a behavioral rule that can be violated by a child.*

Only the obligation and the ban (R1 and R4) can be violated by doing something forbidden: drinking Coke under 12 years of age. R2 and R3 do not express a behavioral restriction and hence cannot be violated.

Participants: 168 students from various disciplines (excluding psychology, mathematics, and philosophy) of the University of Freiburg volunteered for the study and were paid for participating. They were untrained in logic and had no prior experience with selection tasks.

Table 1: (a) *Selection tasks*: Frequency of 'P'/'*not*-Q' responses depending on rule version (n each 28). (b) *Rule evaluation tasks*: Frequency of selecting a rule as one that can be violated ($n = 28$, multiple selections possible).

Rule	(a)	(b)
R1 '*If* P, *then must* Q.'	27	24
R2 '*If* Q, *then may* P.'	27	9
R3 '*If* Q, *then need-not not*-P.'	22	3
R4 '*If* P, *then must-not not*-Q.'	26	25
R5 None	26	

Design and Procedure: The participants were randomly assigned to one of six groups ($n = 28$). Each received one of the six tasks together with other tasks analyzed elsewhere (Beller & Spada, 2000). The treatment was administered in small groups. Each participant received a booklet with a general instruction on the first page and the various tasks each on a new page. Each booklet presented the tasks in a new random order. The order of the conditionals in the rule evaluation task and of the "cards" in the selection tasks was also determined randomly for each participant.

Results: The results of the *selection tasks* are shown in Table 1(a). As predicted, changing the conditional had only a marginal effect ($\chi^2_{(4,\, n\,=\,140)} = 7.84$; $p = 0.10$). Most participants ($m = 91.4$ %) selected the predicted cards 'P'/'*not* Q', even in the task without an explicit conditional. Table 1(b) shows the frequency of selecting each conditional as a rule that can be violated (*rule evaluation task*). Summed up over individual combinations, the predicted conditionals R1 and R4 were selected 49 times while R2 and R3 were selected 12 times (80.3 % vs. 19.7 %; $p < 0.01$, based on the binomial distribution with $n = 61$ and $r = 1/2$). The combination of R1 and R4 was selected by 16 participants.

Summary and Discussion

The experimental results show how violation checking is affected task-specifically by the possibility of constructing normative models. (1) Given the possibility to construct a norm, as in the selection tasks, persons rely on this *norm* and appear to be insensitive to the *form* of the conditional rule describing the norm. This replicates the effect of changing the rule from 'standard' to 'switched' and extends it to other rules. The rule-free version demonstrates that an enriched deontic context (as it is used in many other deontic selection tasks as well) is sufficient to elicit this effect. Since the participants do not regard the conditional rules as relevant premises, their insensitivity to the syntactic form of the rules should not be attributed to illogical reasoning. In fact, persons' answers are in accordance with the logic of social norms. (2) If a task does not allow persons to construct normative models but requires to evaluate deontic conditionals directly, then their answers are indeed quite sensitive to the deontic form. Together,

these findings strongly support the dual source argument (Beller, 1997; Beller & Spada, 2000): in order to understand human deductive reasoning it is necessary to integrate inferences from two sources, namely from the syntactic form of an argument and from conceptual knowledge associated with its content or context.

A mental models notation was used to describe the representation and inferential use of norms (although it is assumed that the deontic principles may be adapted to a mental logic framework as well). The course overview of selection task findings demonstrated how a fine-grained analysis of the domain can guide the interpretation of experimental results. The next step will be to apply DMM theory to findings from other tasks, for example, reformulation tasks or conditional syllogism tasks (e.g., Thompson, 1995), in order to assess its full potential. Two assumptions characterize the proposed representation of norms: the closed world assumption (all norms are known to the reasoner) and the equivalence assumption (concerning the relation between a ban and its conditions). The selection task data are consistent with both. However, they only provide indirect evidence. Reformulation tasks or sufficiency and necessity ratings could prove both assumptions more directly.

Acknowledgements I am grateful to Keith Stenning (University of Edinburgh) as well as to Andrea Bender, Gregory Kuhnmünch, Nikol Rummel, and Hans Spada (Freiburg) for encouraging discussions and helpful comments. This work was supported by a grant from the German Research Foundation (DFG) within the graduate program GK MMI at the University of Freiburg.

References

Beller, S. (1997). *Inhaltseffekte beim logischen Denken*. Lengerich: Pabst.

Beller, S., & Spada, H. (2000). The logic of content effects in propositional reasoning: The case of conditional reasoning with a point of view. Under review.

Chellas, B. F. (1980). *Modal logic: An introduction*. Cambridge: Cambridge University Press.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, *17*, 391-416.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31*, 187-276.

Cummins, D. D. (1996). Evidence for the innateness of deontic reasoning. *Mind & Language, 11*, 160-190.

Dominowski, R. L. (1995). Content effects in Wason's selection task. In S. Newstead, & J. St. B. T. Evans (Eds.), *Perspectives on thinking and reasoning*. Hove: Erlbaum.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning*. Hove: Erlbaum.

Fiddick, L., Cosmides, L., & Tooby, J. (2000). No interpretation without representation: The role of domain-specific representations and inferences in the Wason selection task. *Cognition, 77*, 1-79.

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. *Cognition, 43*, 127-171.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology, 73*, 407-420.

Holyoak, K. J., & Cheng, P. W. (1995). Pragmatic reasoning about human voluntary action: Evidence from Wason's selection task. In S. Newstead, & J. St. B. T. Evans (Eds.), *Perspectives on thinking and reasoning*. Hove: Erlbaum.

Jackson, S. L., & Griggs, R. A. (1990). The elusive pragmatic reasoning schemas effect. *Quarterly Journal of Experimental Psychology, 42A*, 353-373.

Johnson-Laird, P. N. (1978). The meaning of modality. *Cognitive Science, 2*, 17-26.

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge: Cambridge University Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hove: Erlbaum.

Johnson-Laird, P. N., & Byrne, R. M. J. (1992). Modal reasoning, models, and Manktelow and Over. *Cognition*, *43*, 173-182.

Manktelow, K. I., & Over, D. E. (1991). Social roles and utilities in reasoning with deontic conditionals. *Cognition, 39*, 85-105.

Manktelow, K. I., & Over, D. E. (1995). Deontic reasoning. In S. Newstead, & J. St. B. T. Evans (Eds.), *Perspectives on thinking and reasoning*. Hove: Erlbaum.

Noveck, I. A., & O'Brien, D. P. (1996). To what extent do pragmatic reasoning schemas affect performance on Wason's selection task? *The Quarterly Journal of Experimental Psychology, 49A*, 463-489.

Newstead, S., & Evans, J. St. B. T. (Eds.) (1995). *Perspectives on thinking and reasoning*. Hove: Erlbaum.

Platt, R. D., & Griggs, R. A. (1993). Darwinian algorithms and the Wason selection task: A factorial analysis of social contract selection task problems. *Cognition, 48*, 163-192.

Sperber, D., Cara, F., & Girotto, V. (1995). Relevance theory explains the selection task. *Cognition, 57*, 31-95.

Stenning, K., & van Lambalgen, M. (in press). Semantics as a foundation for psychology: A case study of Wason's selection task. *Journal of Logic, Language and Information*.

Thompson, V. A. (1995). Conditional reasoning: The necessary and sufficient conditions. *Canadian Journal of Experimental Psychology, 49*, 1-58.

Thompson, V. A. (2000). The task-specific nature of domain-general reasoning. *Cognition, 76*, 209-268.

van Duyne, P. C. (1974). Realism and linguistic complexity in reasoning. *British Journal of Psychology, 65*, 59-67.

von Wright, G. H. (1963). *Norm and action: A logical enquiry.* London: Routledge & Kegan Paul.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology.* Harmondsworth: Penguin.

Yachanin, S. A. (1986). Facilitation in Wason's selection task: Content and instructions. *Current Psychological Research and Reviews, 5*, 20-29.

Cue Preference in a Multidimensional Categorization Task

Patricia M. Berretty (berretty@fordham.edu)
Department of Psychology,
Bronx, NY 10458-5198 USA

Abstract

Many natural categories vary along multiple dimensions. The present studies address two main questions underlying categorization with multiple dimensions. First, how well can humans perform in a categorization task consisting of five categories varying along nine continuously valued dimensions? Second, what are the properties of the cues preferred by humans if not all the available cues are used? Remarkably, participants not only learned to distinguish among the five categories, but they also learned to do so using only the relevant dimensions. A satisficing model of categorization was best able to account for participants' responses. In addition, in a cue preference task, the results showed that nearly all participants preferred to use the dimension with the greatest variance when the number of dimensions available was restricted, in accord with predictions made by the satisficing model.

Introduction

Categorization has been studied by many disciplines including psychology and machine learning. In the area of psychology, the psychological processes underlying human categorization have been investigated. One common approach to determining these processes has been to teach humans to learn novel categories based on very simple stimuli that vary along only a few dimensions. In such simple situations, the complex calculations involved in some of the popular models of categorization (e.g., Nosofsky's (1986) generalized context model; Ashby's (Ashby & Gott, 1988; Ashby & Perrin, 1988) decision bound theory) may be psychologically plausible. However, the results of these experiments are then assumed to be generalizable to categories whose members vary along many dimensions. It seems unreasonable to assume that humans are capable of the even more complex calculations required with an increase in category dimensionality. For example, Nosofsky, Palmeri, and McKinley (1994) "question the plausibility of exemplar storage processes and the vast memory resources that they seem to require" (p.53).

Machine learning, on the other hand, has been primarily concerned with developing algorithms based on experts in specific domains (Quinlan, 1986) – although the algorithms themselves tend to be general-purpose algorithms (i.e., the algorithms are intended to apply to any categorization task). These algorithms have been developed using large data sets that vary along many dimensions. Therefore, an important step in such algorithms is determining which dimensions from the set of possible dimensions should be used. However, the different methods used to model this step usually involve complex computations and thus are also not psychologically plausible.

What follows is a brief review of two popular categorization models (exemplar models and decision bound theory), as well as a review of a satisficing model of categorization (categorization by elimination). Next, a multi-dimension, multi-category task is described, including a discussion of how well the above three models can account for human responses in such a task. The paper concludes with a brief discussion on the learning of relevant cues in the multi-dimension, multi-category task.

Review of Models

Exemplar Models

Exemplar models (Brooks, 1978; Estes, 1986; Medin & Schaffer, 1978; Nosofsky, 1986) assume that when presented with a novel object, humans compute the similarity between that object and all exemplars of every category in which the novel object could be placed. In theory, the object is placed into the category with which it is most similar, however most exemplar models assume probability matching. Nosofsky's (1986) generalized context model (GCM) allows for variation in the amount of attention given to different features during categorization (see also Medin & Schaffer, 1978). Therefore, it is possible that different cues will be used in different tasks. However, this attention weight remains the same for the entire stimulus set for each particular categorization task, rather than varying across different objects belonging to the same category. GCM uses a probabilistic response rule based on the Luce-Shepard choice model. The probability of placing stimulus i into category j is computed by summing the similarity between stimulus i and all objects in category j along every dimension and then weighting the summed similarity by the bias to respond with category j. The weighted summed similarity is divided by the sum of the weighted summed similarity of stimulus i to each category. Similarity is usually either an exponential (for separable stimuli) or gaussian (for integral stimuli) function of psychological distance (Shepard, 1964). Psychological

distance is computed by the Minkowski r-metric with the addition of two parameters, c and w_k, where c is the discriminability parameter which takes into account that stimuli will look more distinct as experience is gained and w_k is the attention weight for the kth dimension.

Decision Bound Theory

Decision Bound Theory (or DBT—see Ashby & Gott, 1988) assumes that there is a multidimensional region associated with each category, and therefore, that categories are separated by bounds. DBT uses a deterministic response rule. An object is categorized according to the region of perceptual space in which it lies. The perceptual space is divided into regions by decision bounds. For two categories (A and B) each composed of two dimensions (x and y), an object will be placed into category A if the estimated likelihood ratio is greater than some bias, where the likelihood ratio refers to the ratio between the likelihood that stimulus i comes from category A and the likelihood that stimulus i comes from category B. The parameters of this model are b, a response bias; a mean and variance for each dimension (which are usually absorbed into the bound parameters); correlations between pairs of dimensions; as well as parameters to define the decision bound.

Both of these psychological models categorize by integrating cues and using all the cues available (except in exemplar models if a cue has an attention weight of zero). But the memory requirements of these models do differ. GCM assumes that all exemplars ever encountered are stored and used when categorizing a novel object, while DBT only needs to store the bound-determining parameters of each category. In comparison, the Categorization by Elimination algorithm (described below) typically requires as little memory as DBT but it does not integrate all available cues.

Categorization by Elimination

Categorization by Elimination (CBE) was originally developed to describe people's categorization judgments in an animate motion task (see Blythe, Miller, & Todd, 1996). CBE is closely related to Tversky's (1972) Elimination by Aspects model of choice. CBE is a noncompensatory model of categorization, in that it uses cues in a particular order, and categorization decisions made by earlier cues cannot be altered (or compensated for) by later cues. In CBE, cues are ordered and used according to their probability of success. For the present purpose probability of success is defined as a measure of how accurately a single cue categorizes some set of stimuli (i.e., percent correct). This is calculated by running CBE only using the single cue in question, and seeing how many correct categorizations the algorithm is able to make. (If using the single cue results in CBE being unable to decide between multiple categories for a particular stimulus, as will often be the case, the algorithm chooses one of those categories at random—in this case, probability of success will be related to a cue's discriminatory power.)

CBE assumes that cue values are divided up into bins (either nominal or continuous) which correspond to certain categories. To build up the appropriate bin structures, the relevant cue dimensions to use must be determined ahead of time. At present, complete bin structures are constructed before testing CBE's categorization performance. Bins can be constructed in a variety of ways from the training examples by determining low and high cue value boundaries for each category on each dimension. These boundaries are then used to divide up each dimension into the cue-value ranges that form the bins. Thus, CBE only needs to store two values per category per cue dimension, independent of the number of objects encountered.

Categorization with Multiple Dimensions

The majority of psychological studies of categorization have used simple artificial stimuli (e.g., semicircles in two-dimensional space—Nosofsky, 1986) that vary on only a few (two to four) dimensions[1]. This is in contrast to the more natural high-dimensionality machine learning applications, such as wine tasting (Aeberhard, Coomans, & Devel, 1994) or handwriting recognition (Martin & Pittman, 1991). It remains to be demonstrated how optimal humans can be when categorizing objects using many continuously valued dimensions. In addition, the predominant psychological models of categorization have not addressed the issue of how people can categorize a multidimensional object when they are constrained by limited information.

Berretty and her colleagues (Berretty, Todd, & Martignon, 1999; Berretty, Todd, & Blythe, 1997) have illustrated that it is possible for a satisficing model that does not use all the available cues to categorize objects, to perform comparably to integrative models on natural data sets. The purpose of the first experiment in this paper is to investigate whether such a satisficing model is able to account for human categorization data from a multi-dimensional, multi-category task. In Experiment 1a, humans are trained to learn categories that vary along nine dimensions. The generalized context model, categorization by elimination, and a form of decision bound theory will be tested to determine how well each model fits the participants' responses. The purpose of the second experiment is to determine how well humans

[1] Posner and Keele (1968) have used random dot stimuli to test human classification, however, the number of dimensions is indeterminable.

are able to categorize when information is limited. In addition, Experiment 1b investigates the properties of the dimensions people prefer to use when information is limited.

Participants

Four graduate students from the University of California, Santa Barbara participated in Experiment 1a and 1b. All participants had normal or corrected vision. Each participant was paid $8 per hour.

Method

Design The design consisted of five different categories that vary along nine dimensions, where only three of the dimensions are necessary for accurate categorization. The values for each category were generated from a multivariate normal distribution where variance(dim 1) > variance(dim 3) > variance(dim 2), with the variance for the remaining 6 dimensions equal to the variance along dimension 3. All uni-dimensional rules that best separate two categories with the same mean on the other two relevant dimensions have an accuracy of 90% (i.e., category overlap along each pair of dimensions was set to 10%).

Procedure Participants were told that they were to learn five different categories that were equally represented during each learning session. Participants were instructed that they may or may not need to use all the dimensions available to them. Participants were run over consecutive days until learning curves leveled off. Each day consisted of 20 blocks with 50 trials per block (for a total of 1000 trials per day). Stimulus display was response terminated and corrective feedback was given after every trial. Thus, if a subject responded 'A' to an exemplar from category B, a low tone sounded followed by a 'B' appearing on the screen. In addition, overall percent correct was given after every learning block.

A cue preference task (Experiment 1b) was administered to participants the day after learning ended. The cue preference day began with a practice block in which participants simply categorized 50 stimuli as they had done on previous days. The practice block was followed by twelve blocks, each consisting of 50 trials. Each trial began with the presentation of one of the three relevant dimensions. Participants then made a categorization judgment based on only that one dimension. After making a judgment, participants chose another dimension and then made another categorization judgment. Thus, two judgments were made for the same stimulus. The first judgment was based on only one experimenter-chosen dimension, while the second judgment was based on two dimensions. No feedback was given during the last twelve blocks of the test day.

Stimuli and Materials Stimuli were generated using the GRT Toolbox (Alfonso-Reese, 1995). Values along every dimension were transformed from number of dots per square into actual screen coordinates. Each dimension was represented as a texture in one of nine possible squares on a computer screen. The location of the three relevant dimensions was different for each subject with the constraint that the center square (in a 3x3 grid) will never be one of the relevant dimensions. Stimuli were presented on a SuperMac Technology 17T ColorDisplay driven by a PowerMacintosh G3 running a Psychophysics Toolbox (Brainard, 1997) and low-level VideoToolbox (Pelli, 1997) within MATLAB (The MathWorks, Inc., 1998). Each participant sat 18 inches from the monitor. The height of the center square of the stimuli was constrained such that visual angle was less than 2°.

Results and Discussion

Experiment 1A Learning for three of the four participants reached asymptote after five days, while the fourth participant required six days. Participants 1, 2, and 3 achieved an overall accuracy of approximately 70% by the last day, while Participant 4 only achieved an overall accuracy of approximately 60% on the last day. The optimal percent correct was 81.9%. Participants' responses for the last day (without the first block) were randomly split into two halves (training and testing sets) five times. Each split was constrained to contain approximately the same number of stimuli from each category.

The Categorization by Elimination algorithm, the Deterministic Generalized Context Model (see Ashby & Maddox, 1993), and six versions of Decision Bound Theory were fit to each participant's training set responses. For CBE, low and high values of each bin along each dimension, as well as the cue order, were estimated from the responses in the training set. The parameters estimated for GCM were the sensitivity parameter, an attention weight for each dimension, the bias towards each category, and the gamma parameter (which is a measure of response selection). For fitting the GCM, a Euclidean-Gaussian distance-similarity metric was used (see Maddox & Ashby, 1998).

The six versions of DBT were all Independent Decisions Classifiers, which is a special case of Decision Bound Theory in which each dimension is assumed to be independent of the other dimensions (see Ashby & Gott, 1988; Ashby & Maddox, 1990). This version of DBT was used since the best fitting bound (to separate the categories) is perpendicular to each of the three relevant dimensions. In the versions of the Independent Decisions Classifier tested here, one criterion is placed along one dimension. Two criteria are then placed along a second dimension and four criteria are placed along the third dimension. All

Table 1: AIC Scores for Experiment 1A

	P1 Train	P1 Test	P2 Train	P2 Test	P3 Train	P3 Test	P4 Train	P4 Test
GCM	585.4	633.6	739.42	823.08	647.33	687.14	814.4	835.24
DBT	594.74	638.16	742.63	780.87	645.32	665.22	809.55	824.54
CBE	646.28	643.59	638.32	640.36	624.5	634.86	656.04	646.85

possible combinations of the three relevant dimensions were tested.

As mentioned earlier, all three models were fit to part of the data set (the training set) and the best fitting parameters estimates were obtained. These parameters were then used to determine the models' accuracy on the remaining data (the testing set). A potential problem with multi-parameter models is that these models may be prone to overfit the data. That is, they actually fit the noise present in the data in order to achieve high accuracy. Training the model on a subset of the data and testing the model on the rest of the data may assess a model's "true" performance.

The AIC goodness-of-fit statistic was used to compare the fits of the three models.

$$AIC(M_i) = -2\ln L_i + 2v_i$$

Where $\ln L_I$ refers to the negative log likelihood value for model M_i obtained through maximum likelihood estimation and v_i refers to the number of free parameters in model M_i. The smaller the AIC score, the closer the model is to the "true" model (Ashby, 1992).

Goodness-of-fit values for each participant (averaged over the five training and five testing sets) are shown in Table 1. Each row corresponds to one of the three models while each column refers to each participant's training and testing sets. The generalized context model was best able to account for Participant 1's training and testing data. Categorization by elimination was best able to account for Participant 2, Participant 3, and Participant 4's training and testing data.

Experiment 1B Experiment 1b was designed to answer two questions. First, how well can humans perform in a categorization task when dimensionality is reduced? Second, what are the properties of the dimensions preferred by humans? Obviously, one of the most important features of a cue is how accurate that cue is in categorizing objects when used alone. Another property of cues is the range of values possible, that is, the variance of a cue. It seems reasonable to assume that humans are able to learn the accuracy of various cues and would use those cues that are more accurate. Given this assumption, all three of the relevant dimensions are equally accurate when used alone. However, the question of whether humans prefer to use cues with more or less variance is addressed by having different variances for the three relevant dimensions.

In Experiment 1b (conducted after performance asymptotes) participants were given one dimension and asked for a categorization judgment.[2] Participants then chose a second dimension (from the remaining eight dimensions) and made a categorization judgment based on only those two dimensions. Only the three relevant dimensions for the categorization task were used in Experiment 1b as the first cue presented to the participant. Both high and low values of these dimensions were given to the participants. Dimension values were selected from the categories such that the values were always less than (or greater than) the best fitting criteria values for that dimension (i.e., only dimensional values from nonoverlapping category regions were presented).

The first major result to notice from this experiment is the overall percent correct participants achieved, which is shown in Table 2. The optimal percent correct possible with only two categories is 51.6%. Participant 3 was very close to optimal, while Participants 2 and 4 actually performed better than would be expected. In addition, Participant 4 actually performed better in Experiment 1b than in Experiment 1a!

Table 2: Overall Percent Correct in Experiment 1B

	Participant			
	1	2	3	4
Percent Correct	42.67	55.23	49.83	64.5

The results from Experiment 1b indicate that participants did indeed learn which of the cues in Experiment 1a were relevant. All four participants chose (nearly always, if not always) one of the three relevant dimensions as their second cue in Experiment 1b (see Table 3). This indicates that participants were not using any of the other dimensions during Experiment 1a[3].

[2] Participants were given the first cue to insure that all three of the relevant dimensions would be chosen. If participants were allowed to choose the first cue to use, it is possible that the same cue would be used first for each trial.

[3] This does not rule out the possibility that participants were using other dimensions in Experiment 1a, but preferred to use one of the three relevant dimensions when limited in the number of dimensions available to them. However, verbal

Table 3: Dimension Preference for Participants 1-4

Dimension Presented	Dimension Chosen by Participant 1			
	1	2	3	4-9
1	23	150	25	0
2	188	9	2	0
3	186	11	0	0
	Dimension Chosen by Participant 2			
	1	2	3	4-9
1	9	80	103	1
2	86	3	100	5
3	91	88	7	7
	Dimension Chosen by Participant 3			
	1	2	3	4-9
1	16	162	22	0
2	162	5	27	0
3	186	9	4	0
	Dimension Chosen by Participant 4			
	1	2	3	4-9
1	15	45	134	0
2	113	0	87	0
3	133	59	8	0

According to CBE when dimension 1 is presented, dimension 3 should be chosen and when dimension 2 or 3 is presented, dimension 1 should be chosen. When dimension 1 was presented first two of the participants preferred the dimension with the highest probability of success (dimension 3). When dimension 2 was presented first, three of the participants preferred the dimension with the highest probability of success (dimension1). All four participants preferred the dimension with the highest probability of success (dimension 1) when dimension 3 was presented first.

Overall, the participants generally chose the second dimension in accord with predictions made by CBE.

Learning Relevant Cues

Given the difficulty of the task in Experiment 1a, it is remarkable that the participants were able to learn the relevant cues. As shown above, all four participants chose (nearly always, if not always) the three relevant dimensions as their second cue in Experiment 1b. But how did cue use progress as the participants learned the different categories in Experiment 1a? To answer this question three different versions of MDS were fit to the participants' category confusion matrices from each half of each day in order to determine how many cues were used by each participant for a particular data set. MDS_1 uses only one dimension, MDS_2 uses two dimensions, and MDS_3 uses three dimensions to account for the participants' confusions. A χ^2 analysis was performed on the differences between the fit values for models differing in one dimension. These results are reported in Table 4.

For participant 1, an MDS choice model using two dimensions did fit the responses better than an MDS choice model using only one dimension for day 2. By day 4, an MDS choice model using three dimensions did obtain a significantly higher fit value than an MDS choice model using only two dimensions. These results indicate that participant 1 used only one dimension on day 1, two dimensions on days 2 and 3, and three dimensions on days 4 and 5.[4] Similarly, the MDS analysis indicates that participants 2 and 3 used only one dimension on the first half of day 1, two dimensions on the second half of day 1, and three dimensions after day 1. Participant 4 appeared to use only one dimension on the first half of day 1, two dimensions on days 2 and 3, and three dimensions on days 4 through 6. Taken with the results from Experiment 1b, it appears that participants not only increased over days the number of cues used when categorizing, but also learned the correct (or relevant) cues to use to accurately categorize.

Given a task consisting of many dimensions, it is clear that participants begin by using only one dimension. Additional dimensions are then learned in a sequential fashion. What is remarkable from these data, is that participants learned to use all three dimensions. Dimension 1 had more variance than any of the other eight dimensions while dimension 2 had less variance than any of the other eight dimensions. Therefore, it is not surprising that participants were able to learn these two dimensions (i.e., the two dimensions out of nine that had differing variances). Dimension 3 on the other hand, had the same amount of variance as the six irrelevant dimensions, yet participants learned by the end of the experiment that this dimension was necessary for accurate categorization.

Conclusion

In conclusion, the studies reported here show that humans are able to learn artificial multidimensional categories. It was also shown that people are able to distinguish relevant from irrelevant dimensions in multidimensional categorization tasks. Results from such a task indicate that a satisficing model is best able to account for the participants' responses. In addition, the predictions made by the satisficing model regarding cue preference were shown to be in accord with the cue

protocol collected at the end of the experiment indicated that participants were only using three dimensions during Experiment 1a.

[4] Note, that on the last half of day 5, the increase in parameters used by and MDS choice model with three dimensions did not fit the data significantly better than an MDS choice model with less parameters (i.e., less dimensions).

Table 7: X^2_{diff} Values for Participants 1

	Participant 1		Participant 2		Participant 3		Participant 4	
Day/Half	MDS_1-MDS_2	MDS_2-MDS_3	MDS_1-MDS_2	MDS_2-MDS_3	MDS_1-MDS_2	MDS_2-MDS_3	MDS_1-MDS_2	MDS_2-MDS_3
1/1	8.34	0.08	3.26	0.3	8.62	3.6	1.02	2.84
1/2	6.56	6.76	27.18*	12.3	102.9*	18.84*	35.7*	5.08
2/1	83.3*	13.8	71.28*	18.96*	92.78*	9.94	86.16*	0.64
2/2	140.44*	2.56	69.94*	6.54	136.76*	30.16*	117.28*	3.62
3/1	214.98*	9.42	78.76*	22.04*	183.38*	29.14*	109.98*	0.38
3/2	174*	11.14	98.18*	35.86*	140.16*	21.1*	80.2*	4.8
4/1	244.36*	28.54*	116.86*	37.6*	155.02*	35.3*	74.56*	11.92*
4/2	146.22*	22.7*	149.28*	30.82*	196.44*	33.72*	80.36*	22.78*
5/1	151.78*	23.48*	116.8*	38.18*	113.6*	41.34*	80.48*	30.94*
5/2	201.98*	14.5	147.96*	34.34*	193.02*	39.92*	143.76*	18*
6/1	—	—	—	—	—	—	132.96*	37.92*
6/2	—	—	—	—	—	—	155.54*	33.08*

preferences of the participants. Finally, the new experimental design proposed provides a method for further testing the properties of dimensions (cues) that humans prefer (or are constrained?) to use.

References

Aeberhard, S., Coomans, D., & de Vel, O. (1994). Comparative analysis of statistical pattern recognition methods in high dimensional settings. Pattern Recognition, 27(8), 1065-1077.

Alfonso-Reese, L. A. (1995). General Recognition Theory toolbox for MATLAB. [Macintosh computer software], Santa Barbara, CA.

Ashby, F. G., & Gott, R. (1988). Decision rules in the perception and categorization of multidimensional stimuli. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 33-53.

Ashby, F. G., & Maddox, W. T. (1993). Comparing decision bound and exemplar models of categorization. Perception & Psychophysics, 53(1), 49-70.

Ashby, F. G., & Maddox, W. T. (1990). Integrating information from separable psychological dimensions. Journal of Experimental Psychology: Human Perception and Performance, 16(3), 598-612.

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. Psychological Review, 95(1), 124-150.

Berretty, P. M., Todd, P. M., & Blythe, P. W. (1997). Categorization by elimination: A fast and frugal approach to categorization. In M. G. Shafto & P. Langley (Eds.), Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society. Mahwah, NJ: Erlbaum.

Berretty, P. M., Todd, P. M., and Martignon, L. (1999). Using few cues to choose: Fast and Frugal Categorization. In G. Gigerenzer & P. M. Todd (Eds.), Simple heuristics that make us smart. Oxford University Press.

Brainard, D. H. (1997). The Psychophysics Toolbox, Spatial Vision, 10, 443-446.

Maddox, W. T., & Ashby, F. G. (1998). Selective attention and the formation of linear decision boundaries: Comment on McKinley and Nosofsky (1996). Journal of Experimental Psychology: Human Perception and Performance, 24(1), 301-321.

Martin, G. L., & Pittman, J. A. (1991). Recognizing hand printed letters and digits using backpropagation learning. Neural Computation, 3(2), 258-267.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General, 115(1), 39-57.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General, 115(1), 39-57.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies, Spatial Vision, 10, 437-442.

Posner, M. I.; Keele, S. W. (1968) On the genesis of abstract ideas. Journal of Experimental Psychology, 77(3), 353-363.

Quinlan, J. R. (1993). C4.5: Programs for machine learning. Los Altos: Morgan Kaufmann.

Tversky, A. (1972). Elimination by aspects: A theory of choice. Psychological Review, 79, 281-299.

A Perceptually Driven Dynamical Model of Rhythmic Limb Movement and Bimanual Coordination

Geoffrey P. Bingham (gbingham@indiana.edu)
Department of Psychology and Cognitive Science Program, 1101 E. 10th St.
Indiana University
Bloomington, IN 47405-7007 USA

Abstract

We review the properties of coordinated rhythmic bimanual movements and previous models of those movements. Those models capture the phenomena but they fail to show how the behaviors arise from known components of the perception/ action system and in particular, they do not explicitly represent the known perceptual coupling of the limb movements. We review our own studies on the perception of relative phase and use the results to motivate a new perceptually driven model of bimanual coordination. The new model and its behaviors are described. The model captures both the phenomena of bimanual coordination found in motor studies and the pattern of judgments of mean relative phase and of phase variability found in perception studies.

Introduction

In coordination of rhythmic bimanual movements, relative phase is the relative position of two oscillating limbs within an oscillatory cycle. For people without special skills (e.g. jazz drumming), only two relative phases can be stably produced in free voluntary movement at preferred frequency (Kelso, 1995). They are at 0° and 180°. Other relative phases can be produced on average when people follow metronomes, but the movements exhibit large amounts of phase variability (Tuller & Kelso, 1989). They are unstable. Preferred frequency is near 1 Hz. As frequency is increased beyond preferred frequency, the phase variability increases strongly for movement at 180° relative phase, but not at 0° (Kelso, 1990). If people are given an instruction not to correct if switching occurs, then movement at 180° will switch to movement at 0° when frequency reaches about 3-4 Hz (Kelso, 1984; Kelso, Scholz & Schöner, 1986; Kelso, Schöner, Scholz & Haken,1987). With the switch, the level of phase variability drops. There is no tendency to switch from 0° to 180° under any changes of frequency.

These phenomena have been captured by a dynamical model formulated by Haken, Kelso and Bunz (1985). The HKB model is a first order dynamic written in terms of the relative phase, ϕ, as the state variable.

The equation of motion, which describes the temporal rate of change in ϕ, that is, $\dot{\phi}$, is derived from a potential function, $V(\phi)$, which captures the two stable relative phases as attractors as show in Figure 1. The attractors are wells or local minima in the potential layout. As the dynamic evolves, relative phase is attracted to the bottom of the wells at 0° and 180°. A noise term in the model causes the relative phase to depart stocastically from the bottom of a well. The effect of an increase in frequency is represented by changes in the potential. The well at 180° becomes progressively more shallow so that the stochastic variations in relative phase produce increasingly large departures in relative phase away from 180°. These departures eventually take the relative phase into the well around 0° at which point, the relative phase moves rapidly to 0° with small variation.

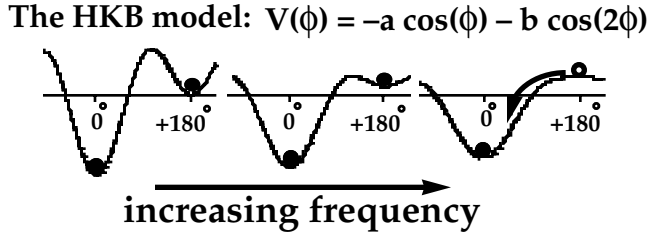


Figure 1. The HKB model. The parameters a and b are varied to model changes in the potential as a function of increases in frequency of movement.

Investigating Phase Perception

We wondered: what is the ultimate origin of the potential function in this model? Why are 0° and 180° the only stable modes and why is 180° less stable than 0° at higher frequencies? To answer these questions, we investigated the perception of relative phase because the bimanual movements are coupled perceptually, not mechanically (Kelso, 1984; 1995). The coupling is haptic when the two limbs are those of a single person. Schmidt, Carello and Turvey (1990) found the same behaviors in a visual coupling of limb movements performed by two different people. Similar results were obtained by Wimmers, Beek, and van Wieringen (1992). To perform these tasks, people must be able to perceive relative phase, if for no other reason, than to comply with the instruction to oscillate at 0° or 180° relative phase.

For reasons discussed at length by Bingham, Zaal, Shull, and Collins (2001), we investigated the visual perception of mean relative phase and of phase variability using both actual human movements (Bingham, Schmidt & Zaal, 1998) and simulations (Bingham, et al., 2001; Zaal, Bingham & Schmidt, 2000) to generate displays of two oscillating balls viewed side on or in depth. Observers judged mean phase or phase variability on a 10 point scale. We found that judgments of phase variability (or of the stability of

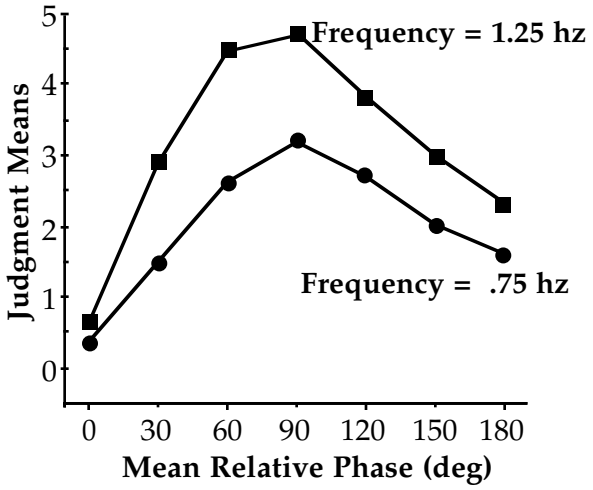


Figure 2. Judgments of phase variability. Mean judgments of phase variability for movements with 0° phase SD (Standard Deviation) and at 7 mean phases from 0° to 180° relative phase. Filled circles: Movement at a freequency of .75 hz. Filled squares: Movement at 1.25 hz.

movement) followed an asymmetric inverted-U function of mean relative phase, even with no phase variability in the movement as shown in Figure 2. Movement at 0° relative phase was judged to be most stable. At 180°, movement was judged to be less stable. At intervening relative phases, movement was judged to be relatively unstable and maximally so at 90°. Levels of phase variability (0°, 5°, 10°, and 15° phase SD) were not discriminated at relative phases other than 0° and 180° because those movements were already judged to be highly variable even with no phase variability. The standard deviations of judgments followed this same asymmetric inverted-U pattern. We found that judgments of mean relative phase varied linearly with actual mean relative phase. However, as phase variability increased, 0° mean phase was increasingly confused with 30° mean phase and likewise, 180° was increasingly confused with 150°. Also, the standard deviations of judgments of mean relative phase followed the same asymmetric inverted-U function found for the means and standard deviations of judgments of phase variability.

Finally, we investigated whether phase perception would vary in a way consistent with the finding in bimanual coordination studies of mode switching from 180° to 0° relative phase when the frequency was sufficiently increased. In addition to mode switching, increases in the frequency of movement yielded increases in phase variability at 180° relative phase but not at 0° relative phase. As shown in Figure 2, Bingham, et al. (in press) found that as frequency increased (even a small amount), movements at all mean relative phases other than 0° were judged to be more variable. This was true in particular at 180° relative phase. Frequency had no effect on judged levels of phase variability at 0° mean phase.

Results from our phase perception studies are all consistent with the findings of the studies on bimanual coordination. The asymmetric inverted-U pattern of the judgments is essentially the same as the potential function of the HKB model. The potential represents the relative stability of coordination or the relative effort of maintaining a given relative phase. The two functions match not only in the inverted-U shape centered around 90° relative phase, but also in the asymmetry between 0° and 180°. 180° is less stable than 0°. This congruence of the movement and perception results supports the hypothesis that the relative stability of bimanual coordination is a function of the stability of phase perception. So, we developed a new model of bimanual coordination in which the role of phase perception is explicit.

Modelling the single oscillator

The HKB model is a first order dynamical model in which relative phase is the state variable. That is, the model describes relative phase behavior directly without reference to the behavior of the individual oscillators. The model was derived from a model formulated by Kay, Kelso, Saltzman and Schöner (1987) that does describe the oscillation of the limbs explicitly. In this latter model, the state variables are the positions and velocities of the two oscillators. To develop this model, Kay, et al. (1987) first modelled the rhythmic behavior of a single limb. In this and a subsequent study (Kay, Saltzman & Kelso, 1991), they showed that human rhythmic limb movments exhibit limit cycle stability, phase resetting, an inverse frequency-amplitude relation, a direct frequency-peak velocity relation, and, in response to perturbation, a rapid return to the limit cycle in a time that was independent of frequency. A dimensionality analysis showed that a second-order dynamic with small amplitude noise is an appropriate model. The presence of a limit cycle meant the model should be nonlinear and a capability for phase resetting entailed an autonomous dynamic. (Note: Phase resetting means that the phase of the oscillator was different after a perturbation than it would have been if not perturbed. An externally driven or non-autonomous oscillator will not phase reset because the external driver enforces its phase which is unaffected by perturbation of the oscillator.) Kay, et al. (1987) captured these properties in a 'hybrid' model that consisted of a linear damped mass-spring with two nonlinear damping (or escapment) terms, one taken from the van der Pol oscillator and the other taken from the Rayleigh oscillator (hence the 'hybrid') yielding:

$$\ddot{x} + b\,\dot{x} + \alpha\,\dot{x}^3 + \gamma\,x^2\,\dot{x} + k\,x = 0 \quad (1)$$

This model was important because it captured the principle dynamical properties exhibited by human rhythmical movements. However, the relation between terms of the model and known components of the human movement system was unclear. The damped mass-spring was suggestive of Feldman's λ-model of limb movement (also known as the equilibrium point or mass-spring model). The λ-model represents a functional combination of known muscle properties and reflexes. Nevertheless, in the hybrid model, the functional realization of the nonlinear damping terms was unknown.

Following a strategy described by Bingham (1988),

Bingham (1995) developed an alternative model to the hybrid model. All of the components of the new model explicitly represented functional components of the perception/action system. The model also incorporated the λ-model, that is, a linear damped mass-spring. However, in this case, the mass-spring was driven by a perceptual term. Limb movements are known to exhibit organizations that are both energetically optimal and stable (e.g. Diedrich & Warren, 1995; Margaria, 1976; McMahon, 1984). Both energy optimality and stability are achieved by driving a damped mass-spring at resonance, that is, with the driver leading the oscillator by 90°. Accordingly, Hatsopoulos and Warren (1996) suggested that this strategy might be used in driving the Feldman mass-spring organization to produce rhythmic limb movements. However, a driver that is explicitly a function of time would yield a nonautonomous dynamic, that is, a dynamic that would not exhibit phase resetting. Bingham (1995) solved this problem by replacing time in the driver by the perceived phase of the oscillator. That is, instead of Fsin(t), the driver is Fsin(ϕ), where ϕ is the phase. Because ϕ (= f[x, dx/dt]) is a (nonlinear) function of the state variables, that is, the position and velocity of the oscillator, the resulting dynamic is autonomous. The perceptually driven model is:

$$\ddot{x} + b\,\dot{x} + k\,x = c\,\sin[\phi] \qquad (2)$$

where

$$\phi = \arctan\left[\frac{\dot{x}_n}{x}\right], \quad \dot{x}_n = \dot{x}/\sqrt{k} \text{ and } c = c(k).$$

The amplitude of the driver is a function of the stiffness. Bingham (1995) showed that this oscillator yields a limit cycle. This is also shown in Figure 3 by rapid return to the limit cycle after a brief perturbing pulse. As also shown, the model exhibits the inverse frequency-amplitude and direct frequency-peak velocity relations as frequency was increased from 1 hz to 6 hz. Finally, the model exhibits a pattern of phase resetting that is similar to that exhibited by the hybrid oscillator. Goldfield, Kay and Warren (1993) found that human infants were able to drive a damped mass-spring at resonance. The system consisted of the infant itself suspended from the spring of a "jolly bouncer" which the infant drove by kicking. This instantiates the model and shows that even infants can use perceived phase to drive such an oscillator at resonance. We hypothesize that all rhythmic limb movements are organized in this way.

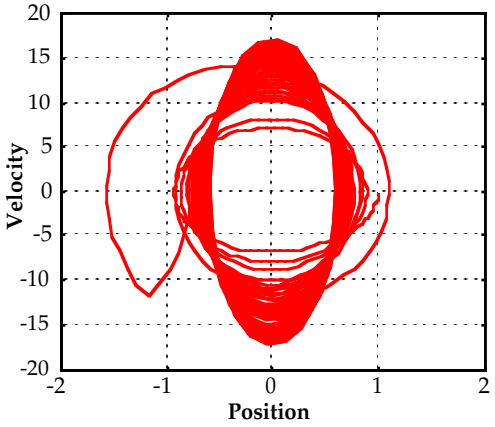


Figure 3. Phase portrait of the single perceptually driven oscillator. Movement starts at 1 hz and increases gradually to 6 hz. Early in the movement while still at 1 hz, the movement was perturbed by a 50ms pulse. Rapid return to the limit cycle within about 1 cycle is shown. Also shown is the decrease in amplitude and the increase in peak velocity that accompanies the increase in frequency.

Once again, the components are the Feldman mass-spring (composed of muscle and reflex properties) and a driver that is a function of the perceived phase of the oscillator.

Modeling Coupled Oscillators

With this model of a single oscillating limb, we were ready to model the coupled system. Kay, et al. (1987) had modeled the coupled system by combining two hybrid oscillators via a nonlinear coupling:

$$\ddot{x}_1 + b\,\dot{x}_1 + \alpha\,\dot{x}_1^3 + \gamma\,x_1^2\,\dot{x}_1 + k\,x_1 = (\dot{x}_1 - \dot{x}_2)[a + b\,(x_1 - x_2)^2]$$

$$\ddot{x}_2 + b\,\dot{x}_2 + \alpha\,\dot{x}_2^3 + \gamma\,x_2^2\,\dot{x}_2 + k\,x_2 = (\dot{x}_2 - \dot{x}_1)[a + b\,(x_2 - x_1)^2] \qquad (3)$$

This model required that people simultaneously perceive the instantaneous velocity difference between the oscillators as well as the instantaneous position differences so that both could be used in the coupling function. This model did yield the two stable modes (namely, 0° and 180° relative phase) at frequencies near 1 hz, and mode switching from 180° to 0° relative phase at frequencies between 3 hz and 4 hz.

We propose an alternative model in which two phase driven oscillators are coupled by driving each oscillator using the perceived phase of the other oscillator multiplied by the sign of the product of the two drivers (P). This sign simply indicates at each instant whether the two oscillators are moving in the same direction (sign = +1) or in opposite directions (sign = –1). The model is:

$$\ddot{x}_1 + b\,\dot{x}_1 + k\,x_1 = c\,\sin(\phi_2)\,P_{ij}$$

$$\ddot{x}_2 + b\,\dot{x}_2 + k\,x_2 = c\,\sin(\phi_1)\,P_{ji} \qquad (4)$$

where

$$P = \mathrm{sgn}(\sin(\phi_1)\,\sin(\phi_2) + \alpha(\dot{x}_i - \dot{x}_j)\,N_t) \qquad (5)$$

P represents the perceived relative phase. As shown in equation (5), the product of the two drivers is incremented by a gaussian noise term with a time constant of 50 ms and a

variance that is proportional to the velocity difference between the oscillators. This noise term reflects known sensitivities to the directions of optical velocities (De Bruyn & Orban, 1988; Snowden & Braddick, 1991) and is motivated by results from phase perception experiments (Collins & Bingham, 2000). This model also yields only two stable modes (at 0° and 180° relative phase) at frequencies near 1 hz, and, as shown in Figure 4, yields mode switching from 180° to 0° relative phase at frequencies between 3 hz and 4 hz. Furthermore, the model predicts our

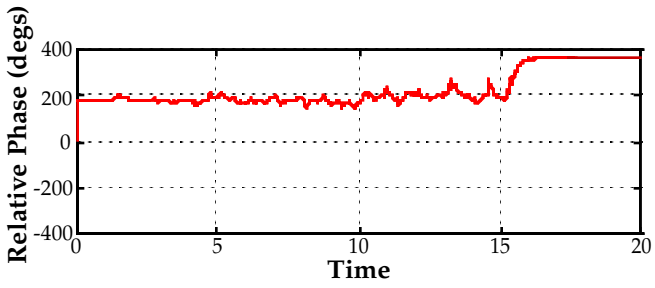


Figure 4. Continuous relative phase from a run of the perceptually coupled model starting at 1 hz and 180° relative phase. Frequency was increased progressively to over 4 hz. Relative phaase became progressively more variable and switched to 360° = 0° at 4 hz. (Note: Frequecy = sqrt(Time+1).)

results for judgments of mean relative phase and of phase variability. (See e.g. Figure 5.) Judged mean phase is produced by integrating P over a moving window of width σ (= 2 s) to yield $\mathbf{P_{JM}}$:

$$\mathbf{P_{JM}} = \frac{\int_{t-\sigma}^{t} P\,dt}{\sigma} \quad (6)$$

Judged phase variability is predicted by integrating $(P - \mathbf{P_{JM}})^2$ over the same window to yield $\mathbf{P_{JV}}$:

$$\mathbf{P_{JV}} = \frac{\int_{t-\sigma}^{t} [P - \mathbf{P_{JM}}]^2 dt}{\sigma} \quad (7)$$

$\mathbf{P_{JM}}$ varies linearly with actual mean phase and $\mathbf{P_{JV}}$ yields an asymmetric inverted-U as a function of actual mean phase.

There are two aspects of the perceptual portions of the model that should be emphasized. First, there are actually two perceptible properties entailed in the model. The two are very closely related, but they are distinct. The first is the phase of a single oscillator. The perception thereof is entailed in the single oscillator model. This is, of course, incorporated into the coupled oscillator model. The second perceptible property is relative phase. This latter property brings us to the second aspect of the model to be noted. This is especially important.

This model is being used to model performance in two different tasks, one is a coordinated movement task and the other is a judgment task. Equation (5) represents the way the perception of relative phase plays a role in the coordinated movement task. This is in terms of the momentary value of P, that is, whether the oscillators are perceived to be moving in the same or in opposite directions at a given moment in time. Equations (6) and (7) represent the way the perception of relative phase plays a role in the judgment tasks. In this case, the behavior of P is assessed (that is, integrated) over some window of time that is large enough to span one or two cycles of movement. So, the two tasks are connected by a single perceptible property, but the way the property is evaluated and used is task-specific.

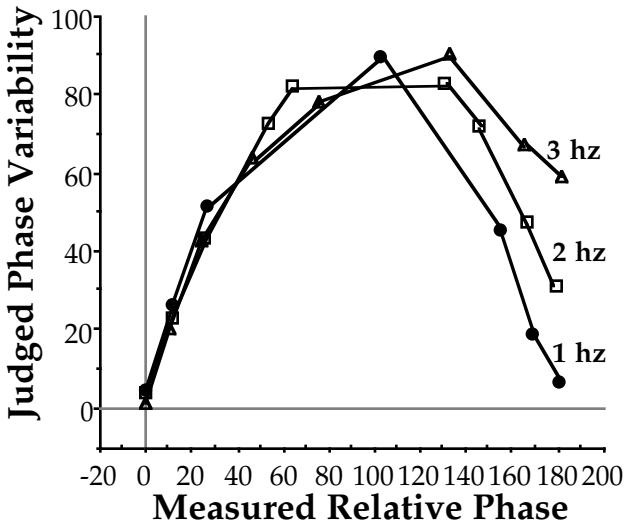


Figure 5. Model predictions of judgments of phase variability at a number of different mean relative phases and at three different frequencies of movement. The model was forced to relative phases other than 0° and 180° to obtain these results.

Conclusions

The model captures both the movement and the perception results. It exhibits the fundamental properties of human rhythmic movements. It builds on the previous task-dynamic modeling results of Kay et al. (1987) and Kay et al. (1991) which revealed fundamental dynamic properties of human movement. Those properties are captured by the new model as they were by previous models. However, unlike the previous models, the new model's components are interpretable in terms of known components of the perception/action system. It explicitly represents the perceptual coupling that is well recognized to be fundamental to the coordination task and the resulting bimanual behaviors. This is important because we can now proceed to investigate the perception component (no less important than the properties of muscle in the Feldman component) to discover the origin of some of the dynamic properties of these perception/action systems. This is an explicit perception/action model.

Finally, although its behaviors are extremely complex,

the model itself is relatively simple and elegant. Two relatively simple equations (4) capture limit cycle stability, phase resetting, inverse frequency-amplitude and direct frequency-peak velocity relationships, the stable modes and mode transitions and the increasing patterns of instability leading up to mode transition. With the addition of two more simple equations (6) and (7) computing a mean and a variance, the model accounts for the results for perceptual judgments of mean relative phase and of phase variability and the ways these vary with the frequency of movement. All this from a model with 5 parameters (k, b, c, α, and σ), four of which are fixed and one, k, is varied to generate variations in frequency of movement. (Note: because c=f(k), c varies with k but once the scaling of c is fixed, this does not represent an extra degree of freedom.) The model is representative of nonlinear dynamics: complex behavior emergent from simple dynamic organization.

Acknowledgments

This research was supported in part by NEI grant # EY11741-01A1 and by NIMH grant # 5T32MH19879-07. The author is grateful for assistance provided by David R. Collins in performing simulations and some of the phase perception studies that have constrained the model. The studies reported herein were reviewed and approved by the Human Subjects Committee at Indiana University. All participants gave their informed consent prior to participation in the experiments.

References

Bingham, G.P. (1988). Task specific devices and the perceptual bottleneck. *Human Movement Science, 7,* 225-264.

Bingham, G.P. (1995). The role of perception in timing: Feedback control in motor programming and task dynamics. In E.Covey, H. Hawkins, T. McMullen & R. Port (Eds.) Neural representation of temporal patterns. New York: Plenum Press.

Bingham, G. P., Schmidt, R. C., Zaal, F. T. J. M. (1998). Visual perception of relative phasing of human limb movements. *Perception & Psychophysics, 61,* 246-258.

Bingham, G.P., Zaal, F.T.J.M., Shull, J.A., Collins, D.R. (2001). The effect of frequency on the perception of the relative phase and phase variability of two oscillating objects. *Experimental Brain Research*, *136*, 543-552.

Collins, D.R. & Bingham, G.P. (2000). How continuous is the perception of relative phase? *InterJournal: Complex Systems*, MS # 381.

De Bruyn, B. & Orban, G.A. (1988). Human velocity and direction discrimination measured with random dot patterns. *Vision Research, 28,* 1323-1335.

Diedrich, F.J. & Warren, W.H. (1995). Why change gaits? Dynamics of the walk-run transition. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 183-202.

Goldfield, G., Kay, B.A. & Warren, W.H. (1993). Infant bouncing: The assembly and tuning of an action system. *Child Develoment, 64,* 1128-1142.

Haken, H., Kelso, J. A. S., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics, 51,* 347-356.

Hatsopoulos, N.G. & Warren, W.H. (1996). Resonance tuning in arm swinging. *Journal of Motor Behavior, 28,* 3-14.

Kay, B.A., Kelso, J.A.S., Saltzman, E.L. & Schöner, G. (1987). Space-time behavior of single and bimanual rhythmical movments: Data and limit cycle model. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 178-192.

Kay, B.A., Saltzman, E.L. & Kelso, J.A.S. (1991). Steady-state and perturbed rhythmical movements: A dynamical analysis. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 183-197.

Kelso, J. A. S. (1984). Phase transitions and critical behavior in human bimanual coordination. *American Journal of Physiology: Regulation, Integration, and Comparative Physiolology, 15,* R1000-R1004.

Kelso, J. A. S. (1990). Phase transitions: Foundations of behavior. In H. Haken and M. Stadler (Eds.), *Synergetics of cognition*. Springer Verlag, Berlin.

Kelso, J. A. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. MIT Press, Cambridge, MA.

Kelso, J. A. S., Scholz, J. P., Schöner, G. (1986). Nonequilibrium phase transitions in coordinated biological motion: Critical fluctuations. *Physics Letters A, 118,* 279-284.

Kelso, J. A. S., Schöner, G., Scholz, J. P., Haken, H. (1987). Phase-locked modes, phase transitions and component oscillators in biological motion. *Physica Scripta, 35,* 79-87.

Margaria, R. (1988). *Biomechanics and energetics of muscular exercise*. Oxford: Clarendon Press.

McMahon, T.A. (1984). *Muscles, reflexes, and locomotion*. Princeton, N.J.: Princeton University Press.

Schmidt, R. C., Carello, C., Turvey, M. T. (1990). Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 227-247.

Snowden, R.J. & Braddick, O.J. (1991). The temporal integration and resolution of velocity signals. *Vision Research, 31,* 907-914.

Tuller, B., Kelso, J. A. S. (1989). Environmentally specified patterens of movement coordination in normal and split-brain subjects. *Experimental Brain Research, 75,* 306-316.

Wimmers, R. H., Beek, P. J., van Wieringen, P. C. W. (1992). Phase transitions in rhythmic tracking movements: A case of unilateral coupling. *Human Movement Science 11,* 217-226.

Zaal, F.T.J.M., Bingham, G.P., Schmidt, R.C. (2000). Visual perception of mean relative phase and phase variability. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 1209-1220.

Inferences About Personal Identity

Sergey Blok (s-blok@northwestern.edu)

George Newman (g-newman@northwestern.edu)

Jennifer Behr (jennifer.behr@yale.edu)

Lance J. Rips (rips@northwestern.edu)

Department of Psychology, Northwestern University,
Evanston, Illinois 60208 USA

Abstract

We investigate the features people use in making inferences about continuity of individual persons. Using a transformation paradigm, we show that people weigh both continuity of the brain and continuity of mental content. Across experiments, we document instances in which participants are more likely to assert individual continuity than continuity of personhood. We discuss these results in terms of a hierarchical view of concepts and philosophical work on personal identity.

Introduction

People are sensitive to the effects that transformations have on membership in basic-level categories (e.g., Gelman & Wellman, 1991; Keil, 1989; Rips, 1989). For example, Rips (1989) asked participants to read stories about creatures of one category (e.g., birds) that came to resemble those of another (e.g., insects). If the transformation was due to accidental factors, participants believed that the creature remained a member of the original category. If the transformation was part of normal maturation, however, participants judged the creature a member of the second category. In general, transformations that alter an object's core properties also tend to change the object's category, while transformations that alter an object's surface or incidental properties tend to preserve its category.

Despite the relatively large number of studies that address questions of category membership continuity, there have been few studies addressing reasoning about individual continuity (see Hall, 1998, Johnson, 1990 and Liittschwager, 1994, for exceptions in the developmental literature). The central question is how we decide that a particular individual – for example, my dog Fido – is the same individual (still Fido) across transformations. This question is potentially different from the one about membership – whether this individual is still a dog.

We investigate here two issues concerning reasoning about individuals. First, we explore the kinds of features people use in judging continuity of identity. Second, we contrast the ways in which people reason about class membership and about identity continuity.

Features of Person Identity

What properties does a person at time t_1 need to share with one at time t_2 in order for that individual to be *the same* at both temporal markers? In making such judgments, people may be *phenomenalists*, relying on continuity of appearance.

In a preliminary experiment, we created stories that varied the type of transformation that a hypothetical target person undergoes. One set of participants -- the Plastic Surgery group -- read a scenario about Jim, a male accountant, who receives plastic surgery to alter his appearance cosmetically to resemble that of Marsha, a female actress. Another set of participants -- the Brain Transplant group -- read a similar story in which Jim's brain is replaced with that of Marsha. After reading the story, both groups supplied judgments of Jim's *identity change* – whether the individual was still Jim or had become Marsha after surgery. Results indicated that a greater proportion of participants in the Brain Transplant group believed Jim's identity had changed than in the Plastic Surgery group (45% and 15%, respectively, $\chi^2(1, 39) = 4.29$, $p < .05$).

These results suggest that changes in appearance are ordinarily not enough to warrant change in identity. The finding parallels earlier studies of natural kinds that show that people tend to reject mere appearance as evidence for category membership when appearance conflicts with deeper properties of the category in question. It has often been suggested that natural kinds' hidden, causally central properties are used for categorization and induction (Ahn, 1998; Gelman & Hirschfeld, 1999), while surface features such as appearance are used for similarity judgments (Keil, 1989; Rips, 1989). But although our results hint that our participants are not folk-phenomenalists, they leave open the question of what criteria they do use to assess personal identity.

The question of criteria for identity is one of the oldest in metaphysics. Writers on the *physicalist* side (e.g., Aristotle, Wiggins) argue that continuity of the body, or, more importantly, the brain, is critical to identity. According to this view, a person P_2 at time t_2 is the same person as P_1 at t_1 if

P_2 has the brain of P_1[1]. Philosophers arguing from a *functionalist* position (e.g., Locke) propose that what matters for identity is not the physical brain, but rather the mental content – the person's unique memories, habits, and personality.

If people are folk-physicalists, then a brain transplant that does not preserve the mental content of the original person should be judged to be as person-preserving as a transplant that retains the original memories. We test this hypothesis in Experiment 1. In Experiment 2 we test the alternative hypothesis that people are folk-functionalists.

Individuals and Hierarchies

A prevalent assumption in the cognitive-psychology literature on categories is that individuals inherit properties of the categories to which they belong. For example, if Fido is a dog, then properties of dogs are true of Fido. (However, see Sloman, 1998, for some exceptions.)

A number of philosophical positions also imply that judgments of identity (Fido vs. Rover) and category membership (dog vs. cat) are related. According to these accounts, criteria of identity for an object (whether phenomenal, physical, or functional) are given by membership in a category to which the object belongs (e.g., Geach, 1962; Gupta, 1980; Wiggins, 1980). If a dog, Fido, is somehow transformed so that it is no longer a dog, it must be true that it is no longer the same individual. In some of these theories (Geach, 1962), different categories to which Fido belongs (e.g., dog vs. pet) yield different criteria of identity, whereas in others (Wiggins, 1980) there can be only one set of criteria. In the experiments that follow, we show that people's reasoning about continuity of identity need not follow *any* obvious category. Instead, participants sometimes rely on distinct sets of features when reasoning about continuity of identity and continuity of category membership.

Experiment 1: Memories and Causality

One goal of Experiment 1 was to determine whether participants perceive continuity of memories as necessary for identity. For the purposes of this paper, we consider "memories" to be unique sets of personal mental representations.

We presented stories that manipulate whether a target person undergoing a brain transplant retains or loses memories in the process. If what matters for identity is continuity of physical parts, such as the brain, then participants should perceive brain transplants that retain memories and those that lose memories as equally conducive to sameness of identity. By contrast, if continuity of memories is essential to continuity of identity, then a brain transplant that retains memories should be more likely to elicit perceptions of sameness than a transplant that does not retain memories.

We also varied whether the memories in question could affect the person's behavior. Because the more essential features of concepts may be those that are causally central (Ahn, 1998), we expected memories that have causal efficacy to be more individual-preserving than memories that could not cause behavior.

One methodological limitation of the preliminary study discussed in the introduction is that participants were never queried about whether the target person was still a member of the same category after the transformation (e.g., whether he was still a person, as opposed to still being Jim). We were therefore not able to determine whether participants' judgments of identity change correlated with their judgments of change in category membership – a correlation that we would expect given the philosophical theories cited earlier. We address this question in Experiment 1 by asking participants to judge the extent to which the post-transformation individual is still a member of his original categories. The category most likely to confer identity on our target individual is the category PERSON itself. However, to determine whether PERSON has special status in this regard, we contrast it with other possible categories, in this case occupation (ACCOUNTANT) and gender (MALE).

Method

Thirty-eight Northwestern University undergraduates read a science fiction story about Jim, a male accountant undergoing a lifesaving brain transplant. Specifically, Jim's brain was transplanted into a robot body. In a between-groups design, we varied whether Jim's memories were the same or different after the transformation. In addition, half the participants in each group answered questions about a situation in which the memories could cause behavior in the robot, and half answered questions about a situation in which the memories could not cause behavior. The full story appears in Figure 1.

The Preserved Memory group read a version of the scenario in which the robot received an *unaltered* version of Jim's brain. The Altered Memory group read a version of the story in which the memories were significantly *altered* during the transformation process.

After reading the story, participants rated their confidence in a number of statements relating to Jim's continuity. Causal efficacy -- whether the memories were able to affect behavior -- was manipulated by varying whether the questions that followed related to events on Monday (when the robot was off) or Wednesday (when the robot was functional). For example, in the Low Causal Efficacy condition, a probe statement read:

On Monday, before the scientists switch it on, the robot is Jim.

0—1—2—3—4—5—6—7—8—9

strongly disagree *strongly agree*

[1] A qualified physicalist position need not require strict sameness of matter, just that there be an unbroken chain of intermediate states between the matter that makes up the body now and the matter that made it up in the past.

The Transplant

Jim is an accountant living in Chicago. One day, he is severely injured in a tragic car accident. His only chance for survival is participation in an advanced medical experiment. Jim agrees.

A team of scientists remove his brain and carefully place it in a highly sophisticated cybernetic body (robot). The robot is powered by electricity. The scientists connect the brain to the robot controls. Though all the right connections between the robot and the brain have been made, the scientists cannot "plug" the robot in because they are waiting for a power adapter they have ordered.

On Monday, the scientists come in to work and the power adapter still has not arrived. While they wait, the scientists scan the brain inside the robot and note that **[THE MEMORIES / NO MEMORIES]** in it are the same as those that were in the brain before the operation.

Finally, on Wednesday, the power adapter arrives and the scientists turn on the robot. The robot appears human-like in its behavior. The robot has senses and can move and talk. Again, the scientists scan the brain inside the robot and find that **[THE MEMORIES / NO MEMORIES]** in it are the same as those that were in the brain before the operation.

Figure 1: Stimulus story for Experiment 1 showing the memory manipulation.

By contrast, in the High Causal Efficacy condition, participants evaluated the statement:

On Wednesday, after the scientists switch it on, the robot is Jim.

0—1—2—3—4—5—6—7—8—9

strongly disagree *strongly agree*

Each participant provided judgments about four kinds of continuity: **individual continuity** ("*… the robot is Jim")*, **personhood continuity** ("*… the robot is a person"*), **gender continuity** ("*… the robot is a male"*), and **occupational continuity** ("*… the robot is an accountant"*). Question order was fully counterbalanced across participants.

Results

Individual Continuity Table 1 summarizes the results of this experiment. As expected, continuity of memory was important for identity. Participants gave higher ratings when the transplanted brain retained the old memories than when it did not (mean ratings on a 0-9 scale were 5.28 and 1.70, respectively; $F(1, 34) = 21.21$, $p < .001$). The capacity of memories to cause behavior also had an effect on continuity. When the robot was described as being "off," continuity judgments were lower than when the robot was "on" (M = 2.15 and 4.83, respectively; $F(1, 34) = 11.83$, $p < .001$). The interaction between Memory and Causal Efficacy was reliable, $F(1, 34) = 5.82$, $p < .05$. Causal efficacy had a larger effect when Jim's memories were retained rather than altered, suggesting that when Jim's old memories were no longer present, it did not matter for identity per se whether the altered memories were able to cause behavior.

	Memories			
	Altered		Preserved	
Continuity	Low CE	High CE	Low CE	High CE
Individual	1.30	2.10	3.00	7.56
Personhood	1.90	3.60	2.00	4.56
Gender	3.80	4.20	2.56	7.00
Occupation	1.30	1.60	2.78	6.78

Table 1: Mean continuity ratings (Experiment 1) as a function of memory continuity and causal efficacy (CE).

Personhood Continuity A similar analysis of personhood ratings showed that, in contrast to identity findings, there was no main effect of Memory, suggesting that continuity of memory was not important for continuity of personhood, $F(1, 34) < 1$. However, there was a main effect of Causal Efficacy, $F(1, 34) = 5.45$, $p < .05$. Participants answering questions about a post-transformation robot in the "on" state were more likely to judge the robot as being a person than those who responded to queries about an "off" robot (M = 4.08 and 1.95, respectively). The interaction between Memory and Causal Efficacy was not reliable, $F(1, 34) < 1$.

Responses to the personhood question were predictive of identity judgments, but the magnitude of the effect was secondary to that of memory continuity. When memory continuity, causal efficacy and personhood were simultaneously regressed onto identity judgments, the standardized regression coefficient for personhood ($\beta = 0.30$) was relatively small compared to that for memory ($\beta = 0.51$). Adjusted R^2 for the overall model was .56.

Gender and Occupation Continuity Gender continuity ("Still a male?") did not fit the pattern of identity judgments. In contrast to identity results, there was no main effect of memories, $F(1,34) < 1$. People's continuity ratings in altered and preserved memory conditions tended towards the middle of the scale (M = 4.00 and 4.7, respectively), indicating that people were not confident about the right way to think about gender continuity. Causal efficacy, however, did play a role in these judgments. When the brain was able to cause behaviors, the object was rated more likely to retain gender than when the brain was unable to cause action (M = 5.60 and 3.18, respectively; $F(1, 34) = 6.62$, $p < .05$). Also, the effect of causal efficacy was greater when memories were retained, as suggested by a reliable interaction between memory and causal efficacy, $F(1, 34) = 4.62$, $p < .05$.

By contrast, occupation ratings ("Still an accountant?") did appear to follow the pattern of the identity ratings. An analysis of variance similar to the one performed on identity continuity judgments revealed parallel effects. We found a main effect of memory continuity such that a brain transplant that preserved memories was also more likely to pre-

serve occupation, $F(1, 34) = 22.35$, $p < .001$. Also, when the brain had an effect on behavior, occupation was more likely to be preserved than if the brain had no effect, $F(1, 34) = 9.33$, $p < .01$. Furthermore, as in the case of the identity ratings, there was an interaction between memory continuity and causal efficacy such that causal efficacy was more important for occupation continuity when memories were preserved, $F(1, 34) = 6.91$, $p < .05$.

Discussion

This experiment provided evidence for the role of memory continuity in perceived identity. Participants who read about a memory-preserving transplant gave higher individual continuity ratings than did participants who read about a transplant that did not preserve memories. This supports the widely held view in the philosophical literature that personal mental representations are central to individual identity. Moreover, our participants granted the highest level of identity continuity to a transplant if these memories had the capacity to cause behavior.

Perhaps the most striking finding is the relative independence between judgments of identity continuity and personhood continuity. People's reasoning about continuity of identity does not appear tightly bound to sameness of membership in normal categories. In fact, participants in some conditions were more likely to agree that the individual was still Jim after the transformation than that he/it was still a person. Specifically, in the condition judged optimal for individual continuity (Preserved Memories, High Causal Efficacy), participants gave a high mean rating of 7.56 when asked if the individual was the same, but a much lower rating of 4.56 when asked if it was still a person, $t(8) = 2.63$, $p < .05$.

We used gender and occupation categories as foils for personhood. As expected, we found only a poor fit between gender continuity and identity. Occupational continuity fared better in terms of reflecting identity judgments, though it seems likely that individual identity was driving occupation identity rather than the reverse. On intuitive grounds, occupational categories are hardly viable as granting identity to individuals. I can cease being a student, without any significant loss of identity. We revisit the issue of individual and occupation identity in Experiment 2.

In general, it seems possible that our participants used different criteria to judge identity, gender, and occupation. Most importantly, people decided about identity and personhood using different criteria. While the critical property of identity appears to be continuity of memories, personhood may depend more heavily on typical properties of persons, such as having a human body and engaging in human behaviors.

Experiment 2 pursues this issue, asking whether perceived identity continuity can be maintained through a transformation that does not preserve any of the physical parts of the original person.

Experiment 2: Necessary Features of Identity

Experiment 1 showed that a brain without the right memories does not guarantee identity of individuals. It is still possible, however, that the brain may be a necessary but not sufficient property of individuals. In this case, memories would have to be transmitted in the physical stuff in which they arose. By contrast, if memories are the "software" that is merely running on the brain "hardware," then it is conceivable that physical brains are not even necessary for identity – any computationally adequate device would do. This is the issue that separates physicalists and functionalists. Will people infer individual continuity even if the original person's memories are "implemented" on a machine that is not the original physical brain?

If the answer is "yes," then we may conclude that people's beliefs about identity are relatively unconstrained, allowing for identity to be preserved through a wide range of fairly extreme transformations. Such a folk-functionalist position is at least intuitively sensible. For example, body cells die and regenerate multiple times throughout the lifespan. It seems odd to consider such physical changes as threatening to personal identity. The competing folk physicalist theory sees brain tissue as at least necessary for identity.

Method

To address the question of physicalism versus functionalism, we modified the brain-transplant scenario from Experiment 1 to include a condition in which the memories in Jim's brain are copied onto a computer designed to control the robot (Computer Copy condition). The story for this condition appears in Figure 2. We also ran a replication of the Brain Transplant scenario from Experiment 1 (Brain Transplant condition) without the passages relating to causal efficacy.

The second factor in the design was whether the memories in the brain (computer) were altered or preserved. This design thus generated four scenarios, which we gave to separate groups of participants. After reading the scenario, participants answered the same set of questions as in Experiment 1. Judgments of individual, personhood, gender, and occupational continuity were made on a 10-point scale.

> Jim is an accountant living in Chicago. One day, he is severely injured in a tragic car accident. His only chance for survival is participation in an advanced medical experiment. Jim agrees.
>
> A team of scientists copy the memories in his brain onto a state-of-the-art computer. The computer is placed in a highly sophisticated cybernetic body (robot). All the right connections between the robot and the computer have been made, and the computer is able to control the robot. The scientists scan the computer and note that [***NONE OF***] the memories in it are the same as those that were in the brain before the operation.
>
> When the scientists turn on the robot, the robot appears to be human-like in its behavior. It has senses and can move and talk.

Figure 2: Stimulus story for the computer copy condition in Experiment 2.

	Memories			
	Altered		Preserved	
Continuity	Computer Copy	Brain Transplant	Computer Copy	Brain Transplant
Individual	0.89	1.53	1.97	5.27
Personhood	1.11	2.83	2.08	2.69
Gender	2.23	4.77	4.34	4.58
Occupation	3.17	0.91	5.78	4.16

Table 2. Mean continuity ratings (Experiment 2) as a function of memory continuity and transplant type.

Questions were presented in two different random orders across participants. Sixty-four Northwestern University undergraduate students took part in the study.

Results

Table 2 presents a summary of the results, which appear to favor a folk-physicalist over a folk-functionalist position.

Individual Continuity Participants in this experiment were more likely to think that the post-transplant individual was still Jim if the transplant included Jim's brain than if it merely included Jim's memories. A Brain Transplant elicited higher continuity ratings than a Computer Copy, $F(1, 56) = 17.95$, $p < .001$. As in Experiment 1, there was also an effect of the memories themselves. Participants who read about a transformation that preserved memories gave higher continuity ratings than those reading about a transformation that altered memories, $F(1, 56) = 26.81$, $p < .001$. Most importantly, however, there was also an interaction between transformation type and memory factors: Preserved memories facilitated continuity to a greater extent when the transformation was a Brain Transplant than when it was a Computer Copy, $F(1, 56) = 8.17$, $p < .01$. There were no reliable effects of question order.

Personhood Continuity As in the case of individual continuity, participants who read the Brain Transplant scenario viewed the robot as more likely to be a person than people who read about a Computer Copy, $F(1, 56) = 4.13$, $p < .05$. However, there was no reliable effect of memory continuity on personhood judgments: Participants were about as likely to think that the robot was a person whether or not the memories were the same as Jim's, $F(1, 56) < 1$.

As this result suggests, personhood continuity ratings did not fully predict judgments of identity continuity. That is, participants were not simply basing their identity judgments (Still Jim?) on whether they believed the object in question is still a person. We tested this claim as we did in the previous study, by running a simultaneous regression with personhood ratings, transplant type, memory continuity, and the interaction between them as predictors of identity responses. The pattern of regression weights closely resembled those in Experiment 1. Memory continuity ($\beta = .46$) was a better predictor of identity judgments than personhood ($\beta = .32$). Adjusted R^2 for the overall model was .54.

Gender and Occupation Continuity A similar analysis of variance on gender continuity ratings revealed no reliable effects. There was a trend for readers of the Brain Transplant story to assert a higher level of gender continuity than participants reading a Computer Copy story ($M = 4.67$ and 3.29, respectively; $F(1, 56) = 3.33$, $p = .07$, n.s.)

Occupational continuity revealed a main effect of transformation type, $F(1, 56) = 8.25$, $p < .01$. It is important to note, however, that the direction of the effect was reversed relative to the identity findings. A Computer Copy was more convincing than a Brain Transplant in allowing Jim to retain the status of an accountant ($M = 4.48$ and 2.53, respectively). As in the previous study, there was also a main effect of memory continuity, $F(1, 56) = 18.71$, $p < .001$. Continuity of memory positively predicted retention of an occupation.

Discussion

These data support the hypothesis that people's naïve construal of individual identity is roughly compatible with folk-physicalism. Continuity of the physical brain had an effect on continuity that went beyond that of functionally equivalent brain content. As in the previous experiment, our data speak against the possibility that this was due entirely to people's beliefs about personhood. While our participants did indicate that a computer copy was less of a person than a Brain Transplant, a regression analysis showed that personhood ratings were not as good a predictor of identity judgments as memory continuity.

We have also replicated the results of Experiment 1 showing that when continuity conditions were optimal (Brain Transplant, Preserved Memories), identity ratings were higher than personhood judgments ($M = 5.27$ and 2.69, respectively; $t(15) = 4.22$, $p < .01$).

Gender and occupation continuity were not good candidates for granting identity to individuals. Both gender and occupation ratings exhibited a poor fit with identity judgments. Occupation continuity actually exhibited a reverse pattern on the criteria people used for continuity judgments. In a number of conditions people were more certain about the continuity of an occupation than they were about individual continuity. For example, in the Computer Copy -- Preserved Memories condition, participants were more likely to say that the object in question is still an accountant than they were to assert that Jim is still in existence (M = 5.78, 1.97, respectively; $t(15) = 3.81$, $p < .01$).

One potential limitation of the current study is that the Computer Copy story was ambiguous as to the fate of Jim's original brain. If our participants assigned identity status to whichever object inherits the original brain, and they believed that Jim's brain survived the accident (even if damaged), then we would expect low continuity ratings in the

Computer Copy condition because Jim's brain is a better candidate for being Jim than the computer containing his memories. This kind of view is proposed by Williams (1973; see Nozick, 1985 for a reply).

While the data can not rule out this possibility, a free response questionnaire administered after the experiment showed that none of our participants explicitly considered Jim's original brain as a factor in their continuity judgments. Furthermore, a related study (in preparation), addressing the question of two possible continuers, showed that people are relatively insensitive to the existence of an identity competitor, preferring to base their judgments on sameness of substance.

Summary and Conclusions

In these studies, we explored the set of features people consider important to personal identity. We showed that when people reason about identity continuity, they take into account continuity of the physical brain and its causally efficacious mental content.[2] People are not phenomenalists, in that appearance is not a necessary feature of an individual's continuity. People are also not unconstrained functionalists, in that they do not assign full continuity if an object only implements a person's unique mental content.

What can a description of a folk theory tell us about the way people form and use concepts? A common assumption is that an individual's identity conditions are given by one or more of the categories to which it belongs. While this provides a convenient way to link categories and individuals, our data show that people do not always use the same sets of characteristics in deciding continuity of an individual and continuity of membership in even its most obvious superordinate category. We have documented instances in which an individual who is viewed as having ceased to be a person is still seen as the same individual.

This finding presents a challenge to the theory that identity conditions are dictated by the superordinate category (or sortal concept) to which that individual belongs. This theory incorrectly predicts that any doubt about proper membership in the category should be reflected in doubt about the individual's survival.

Whether this finding is restricted to reasoning about persons or can be generalized to a wider range of objects remains to be seen. Although relatively minor changes to objects can often cause them to change basic category membership, their individual continuity appears to be much more rigid. Keil (1989) used the example of a coffee pot that was reshaped as a birdfeeder to show that changes in intended function cause shifts in basic-level categorization for artifacts. Despite this change at the category level, however, the object is likely to be judged as the same individual as the one before the transformation.

Finally, reasoning about individuals may turn out to be fundamentally different than reasoning about categories. Individuation often takes into account the history of an object in a way that category membership does not. Whether Jim is still a person after a transformation may depend on whether the causal forces responsible for personhood are still intact. Whether Jim is still Jim, however, may depend on the trajectory of his parts across time.

[2] A strict physicalist position may be question-begging. If being the same individual depends on having the same physical material, how do we decide about whether physical material is the same? As philosophical theories, both the physicalist and functionalist approaches have some important deficiencies, particularly with respect to possible circularity. However, our purpose here is simply to see whether either theory approximates the reasoning of untrained participants. We leave to further research the question of what would happen if our lay-physicalists were confronted with difficulties for their view.

Acknowledgments

This research was supported by NSF grant SES-9907414. The authors would like to thank three anonymous reviewers for their comments.

References

Ahn, W. (1998). Why are different features central for natural kinds and artifacts? *Cognition, 69*, 135-178.

Geach, P. T. (1962). *Reference and generality*. Ithaca, NY: Cornell University Press.

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understanding of the nonobvious. *Cognition, 38,* 213-244.

Gelman, S. A., & Hirschfeld, L. A. (1999). How biological is essentialism? In D. L. Medin & S. Atran (Eds.), *Folkbiology*. Cambridge, MA: MIT Press.

Gupta, A. (1980). *The logic of common nouns*. New Haven, CT: Yale University Press.

Hall, D. G. (1998). Continuity and the persistence of objects. *Cognitive Psychology, 37,* 28-59.

Johnson, C. (1990) If you had my brain, where would I be? Children's understanding of the brain and identity. *Child Development. 61*, 962-972.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Liittschwager, J. C. (1994). *Children's reasoning about identity across transformations*. Doctoral dissertation, Department of Psychology, Stanford University, Stanford CA.

Nozick, R. (1981). *Philosophical Explanations*, Cambridge, MA: Harvard University Press.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. Cambridge, England: Cambridge University Press.

Sloman, S. A. (1998). Categorical inference is not a tree. *Cognitive Psychology, 35,* 1-33.

Wiggins, D. (1980). *Sameness and substance*. Cambridge MA: Harvard University Press.

Williams, Bernard (1973): *Problems of the Self*. Cambridge, England: Cambridge University Press.

Graded lexical activation by pseudowords in cross-modal semantic priming: Spreading of activation, backward priming, or repair?

Jens Bölte (boelte@psy.uni-muenster.de)
Institute of Psychology II, WWU Münster, Fliedner Str. 21
48149 Münster, Germany

Abstract

Spoken word recognition models treat mismatching sensory information differently. Mismatching information deactivates lexical entries according to one type of models. In another type of models, lexical entries are activated reflecting their degree of match with the input. This issue is mainly investigated with cross-modal semantic priming and lexical decision. This paradigm has been suspected of backward priming. I report three cross-modal priming experiments using naming and lexical decision to explore the sensitivity of word recognition to different degrees of match and to investigate the contribution of backward priming. Primes were pseudowords minimally (**baprika*) or maximally (**zaprika*) deviating from a word (*paprika*). A contribution of backward priming processes is likely. A repair of the mispronounced item as suggested by Marslen-Wilson (1993) seems untenable.

Introduction

The tolerance of spoken word recognition to a wide variety of distortions, e.g. misarticulation, phonological processes, or masking by environmental noise poses a serious problem for spoken word recognition models. This problem is tied to the question of what constitutes a match or a mismatch. Certain distortions might not form a mismatch at a lexical level. A system intolerant even to minor mismatches seems unlikely given highly prevalent variation in articulation.

Effects of Match and Mismatch

An account of mismatch effects has been formulated in Cohort (Gaskell & Marslen-Wilson, 1996; Marslen-Wilson, 1993). Gaskell and Marslen-Wilson (1996) found that primes with a legally place assimilated phoneme, e.g. **leam*[1], are as effective as unaltered primes, e.g. *lean*, in cross-modal repetition priming. They suggest that only non-redundant, distinctive, and marked phonological information is coded in the lexicon. For instance, place of articulation is unspecified for coronals: [t],[d],[n],[s],[z]. Therefore, they can assimilate to different places of articulation and still do not form a mismatch at a lexical level (see Coenen, Zwitserlood, & Bölte (in press) for different observations).

Ignoring phonological underspecification, Connine, Blasko, and Titone (1993) found that minimally mismatching primes, e.g. **zervice*, accelerated lexical decision to visual targets, e.g. *tennis*. Maximally deviating primes (e.g. **gervice*, > one phonemic feature) primed if they were made from rare words. Priming effects were graded by degree of mismatch. Connine et al. (1993) concluded that the word recognition system reflects the degree of featural overlap with the input thereby compensating for deviations. These findings are in conflict with observations by Marslen-Wilson (1993; Marslen-Wilson & Zwitserlood, 1989).

McClelland and Elman (1986) claim that Trace can compensate mismatching information. In Trace, activation spreads from a feature level to a phoneme level, and from there to a word level. Activation also spreads back from a word node to its constituent phonemes at the phoneme level. McClelland and Elman (1986) report that minimal mismatches are recognised, e.g. **bleasant* as *pleasant*. But simulations using the whole Trace lexicon showed that Trace does not recover easily from initial mismatches (Goldman, Frauenfelder, & Content, 1997 cited after Frauenfelder & Peters, 1998).

The pattern of results by Connine et al. (1993) can partly be captured by the Shortlist model (Norris, 1994). Input to this model consists of a string of phonemes. A list of words candidates, the Shortlist, is formed on the basis of bottom-up information. Mismatching information lowers the bottom-up support for lexical candidates, similar to the Cohort model. But Shortlist can overcome this reduction and activate the target to "a relatively high level of activation" (Norris, 1994, pp. 214). Bottom-up inhibition prevents activating the target if the mismatch is too large. Systematic simulations investigating the role of mismatching information are not yet available.

Mismatch Repair

Marslen-Wilson (1993, Marslen-Wilson, Moss, & van Halen, 1996) interpreted the findings obtained by Connine et al. (1993) as a reflection of a repair process. According to this view, items like **zervice* are identified as mispronunciations of *service*. **zervice* activates *service* but to a degree insufficient for word recognition. The best fitting lexical entry is determined in a second pass. The target, e.g. *tennis*, might operate as a cue to the identity of the pseudoword. The repair mechanism is a subperceptual process that does not form a new perceptual experience. Marslen-Wilson et al. (1996) suggest that the influence of the repair mechanism is more evident at long interstimulus intervals (ISI) or long reaction-times (RT > 650 ms) because

[1] Pseudowords are marked by an asterix.

processing time available for repair is increased.

Contribution of Backward Priming

Cross-modal priming that is the presentation of an auditory prime accompanied or followed by a visual target was used in all studies mentioned above. This paradigm supposedly reflects backward and forward priming effects in lexical decision (Koriat, 1981; Tabossi, 1996). Backward priming refers to the observation that an association from target to prime in absence of an association from prime to target facilitates reactions.

The most unlikely mechanism for backward priming is expectancy generation. In expectancy generation, the prime is used to generate a set of potential targets. It is unlikely, that the pseudoword primes are used to generate the appropriate targets. One has to assume that the participants recognised the word the pseudoword had been made of and then they used this word to generate the appropriate target set.

At first, spreading of activation also seems to be an unlikely candidate because it is often assumed that the prime influences target processing before the target has been presented. Activation spreads from the prime to the target. Koriat (1981; Kiger & Glass, 1983) proposed that activation also spreads back from the target to the prime. This reactivation of the prime supports target processing. This mechanism requires that the pseudoword activated a lexical entry. Otherwise, there is no entry which could be reactivated by the target. Spreading of activation is restricted to short time frames and therefore predicts a reduction in priming at longer ISIs.

Backward priming is a post-lexical relatedness-checking mechanism according to Seidenberg, Waters, Sanders and Langer (1984). Only forward priming reflects spreading of activation. Lexical decision is especially prone to backward priming processes because of its sensitivity to post-lexical processes (Balota & Lorch, 1986; Neely, 1991). Participants check whether prime and target are related. If the check confirms this, a word answer is given. Chwilla, Hagoort, and Brown (1998) suggest that backward priming reflects post-lexical integration by a fast semantic matching mechanism.[2]

Peterson and Simpson (1989) found backward priming effects also for naming, but they were reduced relative to effects in lexical decision. The introduction of an 300 ms ISI reduced backward priming effects in naming even further while lexical decision was unaffected (Peterson & Simpson, 1989). The differential sensitivity of naming and lexical decision suggests that (backward) priming effects have a different origin in these tasks. Peterson and Simpson propose that backward priming effects reflect mainly lexical retrieval processes in naming but post-lexical bias in lexical decision.

In sum, naming is less influenced by backward priming than lexical decision. If backward priming effects are operative in naming, they reflect lexical retrieval processes. It is possible that research investigating mismatch effects registered backward priming effects because lexical decision has been used. The contribution of a repair mechanism is unclear.

Experimental Issues

The research reported examines backward priming effects in cross-modal priming situations while manipulating the degree of mismatch. I was interested whether pseudowords mismatching lexical entries to different extents produce priming effects and how processes other than forward spreading of activation contribute to these effects. Similar to experiments mentioned above, degree of mismatch was manipulated in broad phonemic classes (voice, place, or manner) ignoring assumptions proposed by phonological underspecification. The contribution of backward priming or a repair mechanism was determined by manipulating the ISI of prime and target and by using lexical decision and naming as tasks. The latter manipulations make this study different from that of Connine et al. (1993).

Naming was used in Experiment 1 and 2. The target followed immediately at prime offset in Experiment 1. In Experiment 2 the ISI was 300 ms. A lexical decision task (LDT) was employed in Experiment 3. The ISI was 0 ms, 250 ms, or 750 ms.

Materials and Pre-tests

The first two experiments employed the same materials and design. A subset of these materials were used in Experiment 3. A target was combined with five different spoken primes. For instance, the visual target *tomato* was either preceded by an semantically related prime, *paprika*, a pseudoword, **baprika,* that minimally deviates from the semantically related prime, or a pseudoword, **zaprika,* that maximally deviates from the semantically related prime. An unrelated word, *library*, or a pseudoword derived from the unrelated word, **nibrary*, served as controls. Minimal deviation means that the pseudoword's first phoneme differed in voice or place from the corresponding phoneme of the base word. A maximally deviating pseudoword differed in at least two phonemic classes from the base word. There were two within-factors: Prime-Target Relatedness (related vs. unrelated) and Prime Type (word, minimal, or maximal). ISI was varied between Experiment 1 and Experiment 2. In Experiment 3, ISI was a between factor for participants but a within factor for items. The experiments required evaluation of the semantic relatedness of prime and target (see Pre-test 1) and the determination that pseudowords were unequivocally perceived as pseudowords (see Pre-test 2).

Pre-test 1 Potential targets (285 in total) were used

[2] Compound cue mechanisms of semantic priming come to similar predictions (Ratcliff & McKoon, 1988). They will not be considered here.

in a cross-modal priming experiment with a LDT. Primes were either semantically related or unrelated to the target. Targets were distributed over lists such that each target appeared only once per list. On list 1 a target *n* was preceded by its semantically related prime while target *m* was preceded by its control prime. Prime conditions of target *n* and *m* were exchanged on list 2. Pseudowords were added to have no-answers for the LDT. Targets were presented at offset of the auditory prime for 250 ms. RTs were measured from target onset for 1500 ms. All prime-target pairs, 197 in total, which attracted less than 50% errors and showed a priming effect larger than 0 ms were selected for Pre-test 2. Mean error percentage for word targets was 8.8%, 5.4% for the semantically related pairs and 12.0% for the unrelated pairs. The average priming effect was 78 ms ($t1(52) = 13.76$, $p < .001$; $t2(196) = 18.81$, $p < .001$).

Pre-test 2 The goal was to investigate whether a minimal pseudoword was unequivocally perceived as a pseudoword. In case a minimal pseudoword is missed, it is equal to its base word and priming might be amplified (Bölte, 1997). Minimal and unrelated control pseudowords were tested in an auditory LDT experiment along with 197 words serving as fillers. RT was measured as before. It was 941 ms for pseudowords and 865 ms for words. Twenty-one pseudowords attracting more than 20% errors were discarded as items for the following experiments. Further 26 pseudowords exhibiting the next highest error rates were excluded. There were 3% errors for minimal pseudowords and 2.6% errors for control pseudowords in the final item set of 75 prime-target pairs.

Experiment 1: Cross-modal Priming Naming with an ISI of 0 ms

This cross-modal priming experiment aimed at investigating the role of mismatching information on lexical activation. There were 75 word targets (see Table 1). Each word target was combined with either a semantically related word, a minimally deviating pseudoword, a maximally deviating pseudoword, an unrelated word, or an unrelated pseudoword. Unrelated pseudowords were always maximally different. Fillers were added such that each prime-target combination was counterbalanced.

On each trial, the visual target was presented in capital letters for 360 ms at prime offset. Participants (63 in total) were instructed to name the target as fast and as accurately as possible.

Table 2 summarises the results of this experiment. Two participants were excluded because of technical failures. Latencies slower than 250 ms or faster than 1500 ms were discarded from the analyses.

Data were analysed using planned comparisons in form of one-sided paired t-tests. Priming effects were evaluated by comparing RTs of the related conditions with RTs of the unrelated conditions. This comparison of RTs provides a measure of priming. Its magnitude can indicate the degree of lexical activation. Word primes facilitated naming responses by 19 ms ($t1(60) = 3.68$, $p < .001$; $t2(74) = 18.97$, $p < .001$). Minimal (11 ms, $t1(60) = 2.73$, $p = .004$; $t2(74) = 2.43$, $p = .009$) and maximal pseudowords (7 ms $t1(60) = 1.94$; $p = .03$; $t2(74) = 1.72$, $p = .04$) were also effective primes.

Table 1: Primes of the target *tomato* as a function of conditions

		minimal	maximal
prime type	word	pseudoword	
related	paprika	*baprika	*zaprika
unrelated	library	*nibrary	

Word primes were more effective than minimal (7 ms, $t1(60) = 7.27$, $p = .017$; $t2(74) = 6.91$, $p = .048$) or maximal pseudoword primes (10 ms, $t1(60) = 2.69$, $p = .005$; $t2(74) = 2.48$, $p = .008$). There was no significant difference between minimal and maximal pseudoword primes (3 ms, both $t < 1$). Thus, there is no gradation in lexical activation between minimal and maximal pseudoword as one observes with lexical decision. The overall smaller priming effect might prevent to distinguish between minimal and maximal pseudowords.

Table 2: Experiment 1. Mean RTs in ms, sd (in parentheses), and error percentages as a function of conditions.

		minimal	maximal
prime type	word	pseudoword	
related	489 (61)	496 (65)	499 (62)
	1.7%	2.6%	2.5%
unrelated	507 (69)	507 (61)	
	2.8%	2.1%	

Still, this finding replicates the one obtained by Connine et al. (1993). Minimal pseudowords and semantically related words prime naming responses. Similar to Connine et al. (1993), maximal pseudowords also show a priming effect. They are able to bring about semantic priming effects if made from rare words which was also the case in this study. The average base word frequency was 23/1 million (Celex, 1995).

There was a graded effectiveness with word primes being most effective. Minimal and maximal pseudowords were less effective than word primes. No further differentiation was possible.

Experiment 2: Cross-modal Priming Naming with an ISI of 300 ms

This experiment used the same materials, design, and task as was used in Experiment 1. The ISI was 300 ms, however. The intention was to examine the effect of the ISI on semantic priming. It was shown previously that backward priming is reduced with such ISI in naming (Peterson & Simpson, 1989). In contrast, a longer ISI provides the repair process with more time available to

repair the deviating input. If it is operative, priming should increase, or, at least, it should not decrease.

There were 53 participants. One participant with more than 15% errors was excluded from the analyses. Data treatment was the same as before. Table 3 displays RTs, sd (in parentheses) and error percentages.

Table 3: Experiment 2. Mean RTs in ms, sd (in parentheses), and error percentages as a function of conditions.

		minimal	maximal
prime type	word	pseudoword	
related	437 (63)	440 (68)	444 (66)
	3.5%	3.5%	4.4%
unrelated	448 (75)	449 (71)	
	4.7%	2.1%	

There was significant priming with word primes (11 ms, $t1(51) = 1.944$, p = .029, $t2(74) = 1.656$, $p = .051$) and with minimal pseudoword primes in *t1* (9 ms, $t1(51) = 1.744$, $p = .044$; $t2 = 1.391$, $p = .084$). There was no priming for maximal pseudoword primes (both $t < 1$). The conditions did not differ statistically from each other (word – min pw: 3 ms, $t < 1$; word – max pw: 7 ms $t1(51) = 1.519$, $p = .068$, $t2(74) = 1.162$, $p = .125$; min pw – max pw: 4 ms, both $t < 1$).

Semantic priming was reduced relative to Experiment 1 and only word and minimal pseudoword primes (in *t1*) accelerated RTs. This finding does not support the repair mechanism suggested by Marslen-Wilson et al. (1996). If such process was operative, an increase in priming should have been obtained.

The activation the target received via spreading of activation by the prime might have been reduced by "normal" decay when finally the target appeared. This resulted in the reduction of priming effects. The same argument also holds for backward priming via spreading of activation. When the target started to "reactivate" the prime, the prime's activation was already decayed.

The following lexical decision experiment served to determine the contribution of specific backward priming accounts. Backward priming processes other than spreading of activation are less affected by long ISIs in lexical decision than in naming. Spreading of activation suffers in lexical decision from the same reduction as in naming (Neily, 1991; Seidenberg, et al. 1984).

Experiment 3: Cross-modal priming with lexical decision

The rational of this experiment was to investigate the influence of the ISI on semantic priming in a cross-modal lexical decision experiment. The ISI was 0 ms, 250 ms, or 750 ms. There were no maximal pseudoword primes here because semantic priming effects are most reliably obtained with minimal pseudowords (see Connine et al., (1993) or Experiment 2).

If there is no or only minor contribution of backward priming processes other than spreading of activation, then priming should be reduced at longer ISIs. If the priming effects of pseudoword primes are due to a repair mechanism, priming should increase with longer ISIs.

The same targets and primes (words and minimal pseudowords) as before were used. In order to have no answers, pseudoword targets were added. Prime-target pairs were added such that Prime Type and Lexical Status of primes and targets were completely counterbalanced.

ISI was varied between participants. The visual target on which the participants had to perform a typical lexical decision task followed an auditory prime. Thirty-five participants were tested per ISI.

Two participants and two items were excluded because of high error rates (> 15%). See Table 4 for mean RTs, sd (in parentheses), and error percentages.

Table 4: Experiment 3. Mean RTs in ms, sd (in parentheses), and error percentages as a function of conditions.

		lexical status	
ISI	prime type	word	min pw
0	related	598 (84)	591 (70)
		3.2%	.8%
	unrelated	626 (76)	614 (77)
		4.7%	3.0%
250	related	610 (70)	605 (96)
		3.6%	4.2%
	unrelated	652 (96)	625 (88)
		4.7%	4.0%
750	related	607 (105)	614 (108)
		4.1%	1.6%
	unrelated	641 (108)	631 (108)
		6.9%	3.5%

ISI was a between factor in the participant analyses but a within factor in the item analyses. All other factors were within factors. The ANOVA yielded a significant result for the main effect Prime Type ($F1(1,100) = 83.409$, $p < .001$; $F2(1,57) = 26.598$, $p < .001$). Lexical decisions were faster for related prime-target pairs (605 ms) than for unrelated prime-target pairs (632 ms). The main effect Lexical status was also significant $F1(1,100) = 7.437$, $p = .008$; $F2(1,57) = 4.265$, $p = .043$). Word pairs (622 ms) were responded to slower than pseudoword pairs (613 ms). ISI was not significant in F1 ($F < 1$) but in F2 ($F2(2,57) = 17.267$, $p < .001$). RTs were faster when the target followed the prime (608 ms) immediately. The ISI of 250 ms or 750 ms delayed the RT by 15 ms or 16 ms, respectively. There was no significant interaction of ISI with any other factor. Only the interaction of Prime Type (related – unrelated) and Lexical Status (word – pw) ($F1(1,100) = 7.446$, $p = .008$; $F2(1,57) = 4.003$; $p = .050$) was significant. The RT obtained in the control conditions differed from each other (word: 639 ms, pw: 623 ms), while the RT in the related conditions did not differ

from each other (word: 605 ms, pw: 603 ms). The related conditions differed from the control conditions (Tukey HSD = 9.145, p = .05). Put differently, the priming effect of the pw condition (20 ms) was smaller than that of the word condition (34 ms).

There is no reduction or any increase in priming across the three ISIs. Thus, no evidence was obtained for a repair mechanism as suggested by Marslen-Wilson et al. (1996). Rather, the stable priming pattern suggests that backward priming processes other than spreading of activation affect the semantic priming effect with pseudoword primes in LDT.

Discussion

Semantic priming effects for words were obtained at all ISIs in all experiments independent of the task. They probably reflect a combination of forward spreading of activation and fast backward priming mechanism such as semantic matching (Chwilla, et al. 1998). Pseudoword primes behaved differently than word primes. They produced (1) smaller semantic priming effects than word primes, (2) they were sensitive to ISIs manipulations in naming but not in LDT.

Our results confirm the finding by Connine et al. (1993) that lexical entries are activated reflecting their degree of match with the input. Even maximal pseudowords are able to activate lexical entries if they are made from rare words. This supports the assumption that a perfect match is not required under certain conditions, e.g. low frequency (or low neighbourhood density?). The flexibility of the word recognition system towards deviations is greater than suggested in Cohort. It fits better to the assumptions formulated in Trace or Shortlist. This flexibility is astonishing because listeners often have to distinguish between minimal pairs. But the employed pseudowords are different from minimal pairs. They were "unique" by fitting best to one specific word, e.g. *paprika*, and not to several words. A word of a minimal pair also fits best to a specific lexical entry but it partly also fits to the other half of the minimal pair. This activation of two lexical entries might be sufficient to cancel out the flexibility observed for pseudowords. The better fitting lexical entry might inhibit the less well fitting one. It is also conceivable that the difference in activation is sufficient to discriminate between the two entries.

The consequences for the repair mechanism are described next. Then the priming mechanisms for naming and lexical decision are introduced and finally some adaptations for the models are outlined.

Marslen-Wilson and colleagues argued that the pseudoword prime activates a lexical entry to a low degree, but not sufficient for normal word recognition. That is, without the target *tomato* one would never recognise *paprika* given **baprika*. Supposedly, this recovery process takes time and is more evident in longer RTs (> 650 ms) or longer ISIs. ISI did not influence the size of the priming effect in lexical decision and RTs were below this "critical" barrier. Thus, it is unlikely that the repair mechanism as suggested by Marslen-Wilson et al (1996) brought about the priming effects. Especially, the assumption that this process "kicks in" later in time does not fit to the data pattern.

The priming effects observed in naming can be a combination of forward spreading of activation and backward priming. They are reduced at longer ISIs because both, forward spreading of activation and backward priming, require a close temporal vicinity of prime and target (Kiger & Glass, 1983). A target presented before prime processing is complete, guides further processing of the prime and thereby influences the final representation according to this view. Notice, that a guidance of prime processing by a target requires that a pseudoword had activated a lexical entry. The smaller temporal overlap of prime-target processing results in a reduction in priming.

Priming effects were unaffected by ISI in the LDT experiment. Priming processes other than spreading of activation influences lexical decisions. Chwilla et al. (1998) showed that semantic matching contributes to priming effects over a range of ISI (0 ms – 500 ms). De Groot (1984, 1985) also postulates a fact-acting post-lexical meaning integration process. The word recognition system searches for a meaningful relationship whenever encountering words. A meaningful relationship between prime and target results in faster lexical decisions because a meaningful relation biases the participants to respond yes in LDT. Such process possibly hinders decisions to unrelated pairs because a failure to find a relation with the prime might bias a no response. That was exactly obtained. But semantic matching can only "form" a meaningful relationship if a prime activates lexical entries. Their degree of activation depend on the amount of phonological overlap with the prime. A semantic relation does not help to name a target more quickly, however. Apparently, semantic matching influences lexical decision more than naming.

To summarise, the present experiments show that pseudowords are able to bring about semantic priming effects. The priming effects are either a consequences of spreading of activation and fast-acting post lexical integration processes. The proposed backward priming mechanism requires the activation of a lexical entry by the prime. Also, pseudowords activate lexical entries according to their degree of overlap.

The models introduced above rely on spreading of activation for explaining semantic priming effects. Post-lexical integration is not implemented in any of the models. I present three reasons for this shortcoming. First, these models were not intended to cover situations in which a meaningful relationship between two words is formed. However, humans look for meaning most of the time. Second, the models miss capabilities to explain flexible task dependent behaviour. Lexical decision and naming as tasks are treated alike by the models. Task dependent capabilities are needed for

covering the findings. Third, post-lexical processes are often understood as task dependent processes in form of a "strategic" adaptation. As such, these processes do not inform about word recognition but rather about "strategic" adaptations. This conception is adequate if the "strategic" processes are under conscious control. But semantic integration is a fast unconscious process (Chwilla et al., 1998, De Groot, 1984, 1985).

Trace, Shortlist, or Cohort describe how semantically related entries influence word recognition in a spreading of activation manner. A situation in which two semantically lexical entries are activated at the same time and form a meaningful relationship, is not taken care of. Two lexical entries are mostly treated as competitors for word recognition. But word recognition can benefit of two simultaneously activated entries, especially given an imperfect input.

Acknowledgements

Part of this research was supported by a grant, BO 1479 / 3-1, of the German research foundation (DFG, Bonn) awarded to Jens Bölte. Thanks go to Pienie Zwitserlood for her support. Heidrun Bien, Gregor Weldert, and Vanessa Mense were so kind to test the participants.

References

CELEX German database (Release D25) [On-line]. (1995). Available: Nijmegen: Centre for Lexical Information [Producer and Distributor].

Bölte, J. (1997). *The role of mismatching information in spoken word recognition.* Hamburg, Germany: Verlag Dr. Kovač.

Chwilla, D.J., Hagoort, P., & Brown, C. (1998). The mechanism underlying backward priming in a lexical decision task: Spreading of activation versus semantic matching. *The Quarterly Journal of Experimental Psychology, 15,* 531-560.

Coenen, E., Zwitserlood, P. Bölte, J. (in press). Variation and assimilation in German: Consequences for lexical access and representation. *Language and Cognitive Processes.*

Connine, C., Blasko, D.G., & Titone, D. (1993). Do the beginnings of words have a special status in auditory word recognition. *Journal of Memory and Language, 32,* 193-210.

De Groot, A.M.B. (1984). Primed lexical decision: Combined effects of the proportion of related prime – target pairs and the stimulus-onset asynchrony of prime and target. *The Quarterly Journal of Experimental Psychology, 36A,* 253-280.

De Groot, A.M.B. (1985). Word-contexts effects in word naming and lexical decision. *The Quarterly Journal of Experimental Psychology, 37A,* 281-297.

Frauenfelder, U.H. (1996). Computational models of spoken word recognition. In T. Dijkstra & K. de Smedt (Eds.), *Computational psycholinguistics (pp. 114-138).* London: Taylor & Francis.

Frauenfelder, U.H. & Peters. G. (1998). Simulating the time course of spoken word recognition: An analysis of lexical competition in Trace. In J. Grainger & A.M. Jacobs (Eds.), *Localist connectionist approaches to human cognition.* Mahwah, NJ: LEA

Gaskell, M.G. & Marslen-Wilson, W.D. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception & Performance, 22,* 144-158.

Goldmann, J.P. Frauenfelder, U.H. & Content, A. (1996). *Information flow in Trace: Recognising words from mismatching sensory inputs.* Manuscript in preparation.

Kiger, J.L. & Glass, A.L. (1983). The facilitation of lexical decisions by a prime occurring after the target. *Memory & Cognition, 11,* 356-365.

Koriat, A. (1981). Semantic facilitation in lexical decision as a function of prime-target association. *Memory & Cognition, 9,* 587-598.

Marslen-Wilson, W.D. (1993). Issues of process and representation in lexical access. In G. Altmann & R. Shillcock (Eds.) *Cognitive Models of Speech Processing: The second Sperlonga meeting,* (pp 187-210). Hove: LEA.

Marslen-Wilson, W.D., Moss, H.E., & van Halen, S. (1996). Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 1376-1392.

Marslen-Wilson, W.D. & Zwitserlood, P. (1989). Accessing spoken words: On the importance of word onsets. *Journal of Experimental Psychology: Human Perception & Performance, 15,* 576-585.

McClelland, J.L. & Elman, J.L (1986). The Trace model of speech perception. *Cognitive Psychology, 18,* 1-86.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G.W. Humphreys (Eds.) *Basic processes in reading: Visual word recognition* (pp. 264 - 336). Hillsdale, NJ: Lawrence Erlbaum Associates.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52,* 189-234.

Peterson, R.R. & Simpson, G.B. (1989). Effect of backward priming in single-word and sentence contexts. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 15,* 1020-1032.

Ratcliff, R. & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review, 95*, 385-408.

Seidenberg, M.S., Waters, G.S., Sanders, M., & Langer, P. (1984). Pre- and postlexical loci of contextual effects on word recognition. *Memory & Cognition, 12,* 315-328.

Tabossi, P. (1996). Cross-modal semantic priming. In F. Grosjean & U.H. Frauenfelder (Eds.), *A guide to spoken word recognition paradigms, (pp. 569-576).* London: Taylor & Francis.

Understanding Visual Categorization from the Use of Information

Lizann Bonnar (LIZANN@PSY.GLA.AC.UK)
Philippe G. Schyns (PHILIPPE@PSY.GLA.AC.UK)
Frédéric Gosselin (GOSSELIF@PSY.GLA.AC.UK)
Department of Psychology, University of Glasgow, 58 Hillhead Street,
Glasgow, Scotland, G12 8QB

Abstract

We propose an approach that allows a rigorous understanding of the visual categorization and recognition process without asking direct questions about unobservable memory representations. Our approach builds on the selective use of information and a new method (Gosselin & Schyns, 2000, *Bubbles*) to depict and measure what this information is. We examine three face recognition tasks (identity, gender, expressive or not) and establish the information responsible for recognition performance. We compare the human use of information to ideal observers confronted to similar tasks. We finally derive a gradient of probability for the allocation of attention to the different regions of the face.

Introduction

In recent years, most face, object and scene recognition researchers have gathered around a common agenda: to understand the structure of representations in memory. A number of fundamental issues have been articulated, and researchers typically ask questions such as: "Are face, object and scene representations viewpoint-dependent? " (Hill, Schyns & Akamatsu, 1997; Perrett, Oram & Ashbridge, 1998; Troje & Bülthoff, 1996; Tarr & Pinker, 1989; Bülthoff & Edelman, 1992; Simons & Wang, 1998, among many others); "Are these representations holistic (e.g, view-based, Poggio & Edelman, 1990; Tarr & Pinker, 1991; Ullman, 1998), or made of smaller components? (e.g,, geons, Biederman, 1987; Biederman & Cooper, 1991)"; "Are internal representations complete (e.g., Cutzu & Edelman, 1996), or sparse? (Archambault, O'Donnell & Schyns, 1999; Rensink, O'Regan & Clark, 1997), "two- or three-dimensional?" (Liu, Knill & Kersten, 1995), "colored or not?" (Biederman & Ju, 1988; Oliva & Schyns, 2000; Tanaka & Presnell, 1999), "Are they hierarchically organized in memory?" (Jolicoeur, Gluck & Kosslyn, 1984; Rosch, Mervis, Gray, Johnson & Boyes-Braem, 1976), "Is there a fixed entry point into the hierarchy?" (Gosselin & Schyns, in press; Tanaka & Taylor, 1991) "Does expertise modify memory representations?" (Biederman & Shiffrar, 1987; Tanaka & Gauthier, 1998; Schyns & Rodet, 1997) and the entry point to recognition?" (Tanaka & Taylor, 1991); "What is the format of memory representations, and does it change uniformly across the levels of a hierarchy?" (Biederman & Gerhardstein, 1995; Jolicoeur, 1990; Tarr & Bülthoff, 1995).

To address these complex issues, recognition researchers should be equipped with methodologies of a commensurate power; methodologies that can assign the credit of behavioral performance (e.g., viewpoint-dependence, configural effects, color, speed of categorization, point of entry, expertise and so forth) to specific properties of the representations of visual events in memory. However, the relationship between behavior and representations is tenuous, making representational issues the most difficult to approach experimentally.

In this paper, we propose an alternative approach that allows a rigorous understanding of the recognition process, without directly asking questions about unobservable memory representations. Our analysis builds on the *selective use of diagnostic information*, an important but neglected stage of the recognition process. To recognize an object, people selectively use information from its projection on the retina. This information is not available to conscious experience, but the visual system knows what it is, and how to extract it from the visual array. Our approach interrogates the visual system to determine and to depict the information the system uses to recognize stimuli.

The aim of this paper is twofold. At an empirical level, we will use Gosselin and Schyns (2000) *Bubbles* technique to visualize the information used in three face categorization tasks (identity, gender and expressive or not). Faces are a good stimulus for our demonstrations: their compactness enables a tight control of presentation which limits the spatial extent of useful cues; the familiarity of their categorizations simplifies the experimental procedure which does not require prior learning of multiple categories--most people are "natural" face experts (Bruce, 1994). However, the principles developed with faces also apply to the more general cases of objects and scenes.

At a more theoretical level, we will reveal the information used in recognition tasks without asking questions (or even making assumptions) about memory representations. This is nonetheless a powerful approach because the information used encompasses all the visual features that mediate the recognition task at

hand. These features therefore reflect the information required from memory to recognize the stimulus; their extraction from the visual array specifies the job of low-level vision. Shortly put, the features involved in a recognition task bridge between memory and the visual array. Now, show me the features!

Experiment

This experiment was cast as a standard face categorization and recognition experiment. In a between-subjects design, a different subject group resolved one of three possible categorizations (identity, gender, expressive or not) on the same set of ten faces (5 males, 5 females), each displaying two possible expressions (neutral vs. happy). Prior to the experiment, all subjects learned the identity of the ten faces, in order to normalize exposure to the stimuli.

To determine the specific use of face information in each task, we applied Gosselin and Schyns' (2000) *Bubbles* technique. *Bubbles* samples an input space to present as stimuli sparse versions of the faces. Subjects categorize the sparse stimuli and *Bubbles* keeps track of the samples of information that lead to correct and incorrect categorization responses. From this information, we can derive the usage of each region of the input space for the categorization task at hand (see Figure 1). In a nutshell, *Bubbles* performs an exhaustive search in a specified image generation space (here, the image plane x spatial scales), using human recognition responses to determine the diagnostic information.

Methods

Participants.

Participants were forty-five paid University of Glasgow students, with normal, or corrected to normal vision. Each participant was randomly assigned to one of three possible experimental groups (IDENTITY; male vs. female, GENDER; expressive or not, EXNEX) with the constraint that the number of participants be equal in each group.

Stimuli.

All experiments reported in this paper ran on a Macintosh G4 using a program written with the Psychophysics Toolbox for Matlab (Brainard, 1997; Pelli, 1997). Stimuli were computed from the greyscale faces of Schyns and Oliva (1999) (5 males, 5 females each of whom displayed two different expressions, neutral and happy, with normalized hairstyle, global orientation and lighting).

To search for diagnostic information, we used Gosselin and Schyns' (2000) *Bubbles* technique applied to an image generation space composed of three dimensions (the standard X and Y axes of the image plane, plus a third Z axis representing 6 bands of spatial frequencies of one octave each). Figure 1 illustrates the stimulus generation process.

To compute each stimulus, we first decomposed an original face into 6 bands of spatial frequencies of one octave each—at 2.81, 5.62, 11.25, 22.5, 45 and 90 cycles per face, from coarse to fine, respectively (computations were made with the Matlab Pyramid Toolbox, Simoncelli, 1997). The coarsest band served as a constant background, as a prior study revealed that it does not contain face identification information.

The face represented at each band was then partly revealed by a mid-grey mask punctured by a number of randomly located Gaussian windows (henceforth called "bubbles"). The size of the Gaussian differed for each frequency band, to normalize to 3 the number of cycles per face that any bubble could reveal (standard deviations of bubbles were 2.15, 1.08, .54, .27, and .13 cycles/deg of visual angle, from coarse to fine scales). Since the size of the bubbles decreases from coarse to fine scales, we multiplied the number of bubbles at each scale to normalize the average area of the face revealed at each scale.

To generate an experimental stimulus, we simply added the information revealed at each scale. The total subspace revealed by the bubbles (and therefore the number of bubbles per scale) was adjusted to maintain categorization of the sparse faces at a 75% correct criterion.

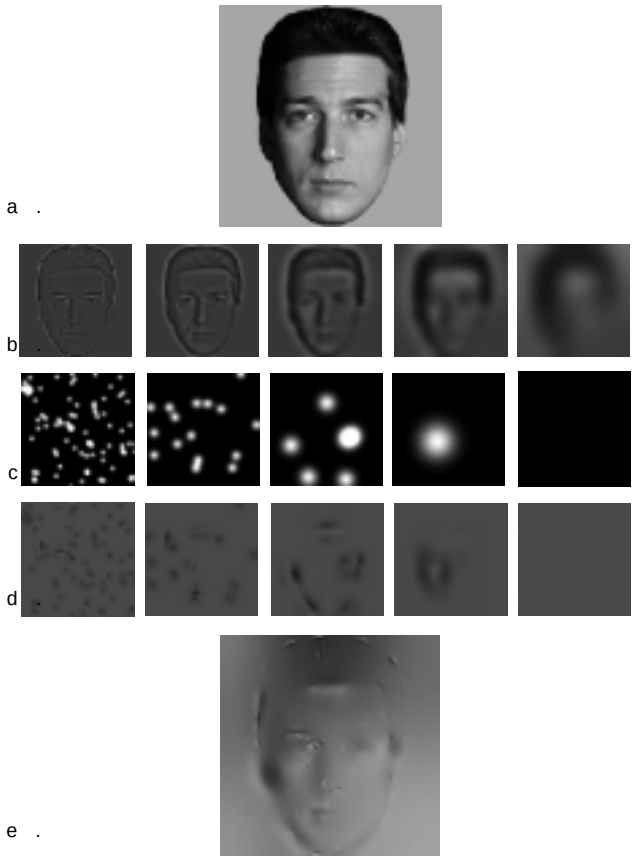


Figure 1 illustrates the application of Bubbles to the 3D space composed of a 2D face in Experiment 2. Pictures in (b) represent five different scales of (a); (c) illustrate the bubbles applied to each scale; (d) are the revealed information of (b) by the bubbles of (c). Note that on this trial there is no revealed information at the fifth scale. By integrating the pictures in (d) we obtain (e), a stimulus subjects actually saw.

Procedure

Prior to experimentation, all participants learned to criterion (perfect identification of all faces twice in a row) the gender, expression and the name attached to each face from printed pictures with corresponding name at the bottom. Each participant was then randomly assigned to one of the three different categorization tasks. In IDENTITY, participants had to determine the identity of each face stimulus. In the GENDER task, participants were instructed to decide whether the stimulus was male or female. In EXNEX, participants had to judge whether the face was expressive or not. Thus, each group performed different categorizations on the same stimulus set.

In a trial, one sparse face computed as just described appeared on the screen. To respond, participants pressed labelled computer-keyboard keys. No feedback was provided. The experiment comprised two sessions of 500 trials (25 presentations of the 20 faces), but we only used the data of the last 500 trials, when subjects were really familiar with the faces and experimental procedure. A chinrest was used to maintain subjects at a constant viewing distance (of 100 cm). Stimuli subtended 5.72 x 5.72 deg of visual angle on the screen.

Results

On average, a total of 33, 20 and 15 bubbles were needed for subjects to reach the 75% performance criterion in the identity, gender and expressive or not task, respectively. Remember that these bubbles resided at different scales of the same stimulus, and were randomly distributed within each scale. Thus, *Bubbles* performs a random search of the input space that is exhaustive after many trials.

Following Gosselin and Schyns' (2000) methodology, we used subjects responses to determine which stimulus information was, and was not diagnostic. The correct categorization of one sparse stimulus indicates that the information revealed in the bubbles was sufficient for its categorization. When this happened, we added the mask of bubbles to a CorrectPlane, for each scale—henceforth, CorrectPlane(scale), for scale = 1 to 5. We also added these masks to a TotalPlane(scale), for each scale. Across trials, TotalPlane(scale) represents the addition of all masks leading to a correct categorization *and* a miscategorization.

From CorrectPlane(scale) and TotalPlane(scale), we can compute for each subject the diagnositicity of each region of the input space with ProportionPlane(scale) = CorrectPlane(scale)/TotalPlane(scale). For each scale, the ProportionPlane(scale) is the ratio of the number of times a specific region of the input space has led to a successful categorization over the number of times this region has been presented. Across subjects, the averaged ProportionPlane(scale) weighs the importance of the regions of each scale for the categorization task at hand (Gosselin & Schyns, 2000). If all regions had equal diagnosticity, ProportionPlane(scale) would be uniformly grey. That is, the probability that any randomly chosen bubble of information led to a correct categorization of the input would be equal to the performance criterion—here, .75. By the same reasoning, whiter regions are significantly above the performance criterion, and therefore more diagnostic of these tasks.

To compute the significance of diagnostic regions, a confidence interval is built around the mean of the ProportionPlane(scale), for each proportion ($p < .01$). To depict the complex interaction between categorization tasks, spatial scales and use of information, we can visualize the *effective stimulus* of each task (see Figure 2). The effective stimulus is a concrete image of the information the visual system uses in each task. It is obtained by multiplying the face information in Figure 2 with the diagnostic masks.

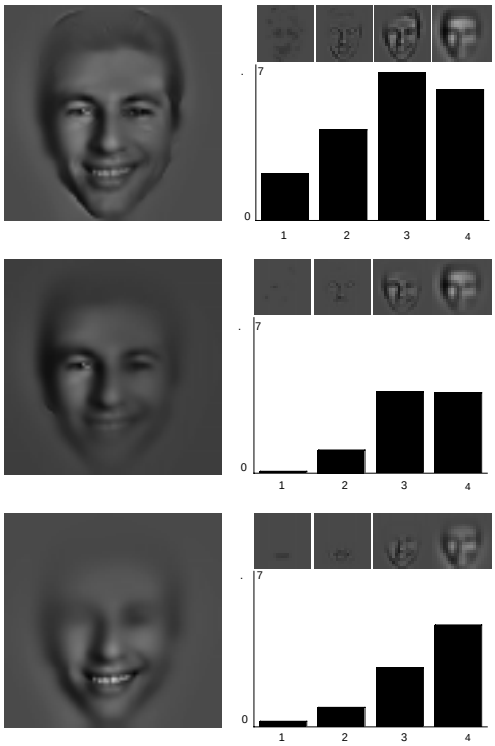


Figure 2. (a) The larger face depicts the effective face stimulus for the identity task. The smaller pictures illustrate the diagnostic information used to resolve the identity task at each independent scale from fine to coarse, respectively. The coarsest scale is not depicted as it contains no meaningful information. The bar chart provides a quantitative illustration of the proportion of the face area used to resolve the task at each scale. Figures (b) and (c) follow the same format as figure (a) illustrating the potent face for the gender task and expressive or not task respectively, the diagnostic information for each task at each scale and a quantitative account of the use of information is illustrated in the bar charts.

Discussion

Use of scale information between categorization tasks. Figure 2 presents a comparison of the relative use of scale information across tasks. From top to bottom, the large face pictures depict the information used in identity, gender and expressive or not. The figure reveals that the use of diagnostic information differs across categorization tasks, and scales. For example, whereas the mouth is well-defined at all scales in the identity and expressive tasks it is neglected at the finest scales in the gender task. In a related vein, the eyes are both represented at all scales in identity, but only one of them is well represented in gender, and both are neglected in expressive. The chin is well defined in identity, but not in expressive and gender. Compared to the mouth and the eyes, the nose is less well defined in all tasks.

To quantify the use of spatial scales across tasks, we computed the diagnostic areas revealed at each scale over the total area covered by the face in the image plane. The histograms in Figure 2 plot the use of diagnostic information across spatial scales--1 means finest, and 4 coarsest scale. The small face pictures corresponding to each scale illustrate what this face information is. The pictures reveal that the use of fine scale information (labelled 1 in the histograms, and depicted in the leftmost small picture) is most differentiated across the three tasks. In identity, it depicts the eyes, the mouth and the chin, whereas in gender it is only used for the left side eye, and the mouth in expressive. In contrast to the finest scale, the coarsest scale (i.e., the fourth scale) is much less differentiated, revealing only a holistic representation of the face features. This forms a skeleton that is progressively distinguished and fleshed out with increasing spatial resolution (see the progression of face information from coarse to fine in the small pictures of Figure 2, from right to left.) The asymmetry in extracting diagnostic information to resolve the gender task is consistent with studies showing that there is a right-hemisphere bias (the left-side of the image) in processing various facial attributes, including gender (Burt & Perrett, 1997).

Turning to the relative use of scales within each task, there is a clear advantage for the third scale in identity, corresponding to face information comprised between 11.25 and 22.5 cycles per face. This is consistent with the face recognition literature where the best scale for face recognition is between 8 and 32 cycles per face, depending on authors (see Morrison & Schyns, in press, for a review). Note, however, that our analysis is more refined because not only can we specify what the best scale is, but also where this information is located in the image plane. In contrast, the best scale for expressive or not (here, the discrimination between neutral and happy) is information comprised between 5.62 and 11.25 cycles per face (the fourth scale). This is in line with Jenkins et al. (1997) and Bayer, Scwartz & Pelli, (1998) who also found that the detection of the happy expression was most resilient to changes in viewing distances (i.e., with information corresponding to coarser scales). For gender, scales 3 and 4 were most used, and across task, there appears to be a bias for face information comprised between 5.62 and 22.5 cycles per face (the coarser scales) when information was available from the entire scale spectrum. At this stage, it is worth pointing out that the self-calibration property of Gosselin and Schyns' (2000) technique ensures that if subjects required only information from the finest scale to resolve the tasks, they would not reach the performance criterion of 75% and the average number of bubbles would increase at each scale, until they displayed enough information at the finest scale to reach criterion. In other words, the reported biases for the coarser scales do not arise from the technique, which is unbiased, but from the biases of observers who use information in categorization tasks.

Ideal Observers. In *Bubbles*, the observer determines the informative subset of a randomly, and sparsely sampled search space. To highlight this unique property, we here contrast human and ideal observers (Tjan, Braje, Legge & Kersten, 1987). The ideal observer will provide a benchmark of the information available in the stimulus set to resolve each task. We have biased the ideal to capture all the regions of the image that have highest local variance between the considered categories (identity, male vs. female, and neutral vs. expressive). This ideal considers the stimuli as images (not as faces composed of eyes, a nose and a mouth, as humans do). The ideal might not necessarily be sensitive to the regions that humans find most useful (the diagnostic regions), but to the information that is mostly available in the data set for the task at hand. We constructed a different ideal observer for the tasks of identity, gender, and expressive or not and submitted them to *Bubbles*, using the same parameters as those of our experiment with humans. Here, however, the number of bubbles remained constant (equal to the average required in each task), and we added to the face

stimuli a varying percentage of white noise to maintain categorization performance at 75% correct. In a Winner-Take-All algorithm, the ideal matched the information revealed in the bubbles with the same bubbles applied to the 32 memorized face pictures. The identity, gender or expressive or not categorization response of the ideal was the best matched picture. We then computed the ProportionPlane(scale) and DiagnosticPlane(scale), as explained earlier, to derive the effective face of each categorization task (see Figure 3). A comparison between the human and the ideal effective faces reveal only a partial correlation of use of information. This indicates that the highest variations of information in the image were not necessarily used by humans, who instead focused on the diagnostic face information. It further stresses that *Bubbles* is a human, partially efficient, not a formal, optimally efficient, feature extraction algorithm (Gosselin & Schyns, 2000).

Figure 3. The effective face stimulus of the Ideal Observer for each categorization task, identity, gender and expressive or not, respectively.

Deriving a two-dimensional map of attention. So far, we have examined the use of information across the different spatial scales of a face. We can now derive a precise measure of the diagnosticity of each image locations for the different face categorizations. Remember that the DiagnosticPlane(scale) represent a with value of 1 the presence of diagnostic information at all image locations. To measure the gradient of probability of finding diagnostic information at any image location, we simply multiply the normalized probability of using a scale with the DiagnosticPlane of this scale, and add together all the DiagnosticPlane(scale). When diagnostic information was present (vs. absent) at all scales for this image location, it has a probability of 1 (vs. 0). Figure 4 renders with a grey scale the gradient of probability (white = 1, black = 0) of finding diagnostic information at any location of the image in identity, gender, and expressive or not. If the attention is alloctated (or eye movements are guided) to the most relevant image locations in a task, the maps of Figure 4 have a predictive value. For example, Figure 2 reveals that the regions of the eyes and the mouth are diagnostic across the entire scale spectrum, and so these locations have highest probability in Figure 4. From the seminal work of Yarbus (1965), studies in eye movements have consistently demonstrated that the eyes and the mouth were mostly scanned in face identification tasks.

Figure 4. The 2D attentional maps for each categorization task, identity, gender and expressive or not, respectively.

Concluding Remarks

Our goal was to address the problem of recognition without directly asking questions about internal representations. Our analysis established how three face categorization tasks selectively used information from a three-dimensional input space (the two-dimensional image plane x spatial scales). From this selective use, we derived a gradient of probability of locating diagnostic information in the image plane. A rational categorizer should selectively allocate its attention to the regions of the image that maximize this probability thus minimizing the uncertainty of locating diagnostic information, see Figure 4.

Acknowledgements

This research was supported by ESRC grant R000223179.

References

Archambault, A., O'Donnell, C., & Schyns, P.G. (1999). Blind to object changes: When learning one object at different levels of categorization modifies its perception. *Psychological Science*, **10**, 249-255.

Bayer, H.M., Schwartz, O., & Pelli, D. (1998). Recognizing facial expressions efficiently. *IOVS*, **39**, S172.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, **94**, 115-147.

Biederman, I., Shiffrar, M.M. (1987). Sexing day-old chicks: a case study and expert systems analysis of a difficult perceptual leaning task. *Journal of Experimental Psychology: Learning, Memory & Cognition,* **13**, 640-645.

Biederman, I., & Cooper, E.E. (1991). Priming contour-deleted images: Evidence for intermediate representations in visual object recognition. *Cognitive Psychology*, **23,** 393-419.

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*, **20**, 38-64.

Biederman, I. & Gerhardstein, P.C. (1995). Viewpoint-dependent mechanisms in visual object recognition: a critical analysis. *Journal of Experimental Psychology: Human Perception and Performance,* **21**, 1506-1514.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, **10**, 433-436.

Bruce, V. (1994). What the human face tells the human mind: Some challenges for the robot-human interface. *Advanced Robotics*, **8**, 341-355.

Bülthoff, H.H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view theory of object recognition. *Proceedings of the National Academy of Science USA*, **89**, 60-64.

Burt, D.M. & Perrett, D.I. (1997). Perceptual asymmetries in judgements of facial attractiveness, age, gender, speech and expression. *Neuropsychologia,* **35**, 685-693.

Cutzu, F., & Edelman, S. (1996). Faithful representations of similarities among three-dimensional shapes in human vision. *Proceedings of the National Academy of Science*, **93**, 12046-12050.

Gosselin, F. & Schyns, P.G. (2000). Bubbles: A new technique to reveal the use of visual information in recognition tasks. Submitted for publication.

Gosselin, F & Schyns, P.G. (in press). Why do we SLIP to the basic-level? Computational constraints and their implementation. *Psychological Review.*

Hill, H., Schyns, P.G., & Akamatsu, S. (1997). Information and viewpoint dependence in face recognition. *Cognition.* **62**, 201-222.

Jenkins, J., Craven, B., Bruce, V., & Akamatsu, S. (1997). Methods for detecting social signals from the face. Technical Report of IECE, HIP96-39. The Institute of Electronics, Information and Communication Engineers, Japan.

Jolicoeur, P. (1990). Identification of disoriented objects: A dual-systems theory. *Mind and Language*, **5**, 387-410.

Jolicoeur, P., Gluck, M., & Kosslyn, S.M. (1984). Pictures and names: Making the connexion. *Cognitive Psychology*, **19**, 31-53.

Liu, Z., Knill, D.C., & Kersten, D. (1995). Object classification for human and ideal observers. *Vision Research*, **35**, 549-568.

Morrisson, D. & Schyns, P.G. (in press). Usage of spatial scales for the categorization of faces, object and scenes. *Psychological Bulletin and Review.*

Oliva, A. & Schyns, P.G. (2000). Colored diagnostic blobs mediate scene recognition. *Cognitive Psychology*, **41**, 176-210.

Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, **10**, 437-442.

Perrett, D.I., Oram, M.W., & Ashbridge, E. (1998). Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformation. *Cognition*, **67**, 111-145.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, **343**, 263-266.

Rensink, R.A., O'Regan, J.K., & Clark, J.J. (1997). To see or not to see: the need for attention to perceive changes in scenes. *Psychological Science,* **8**, 368-373.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D. & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, **8**, 382-439.

Schyns, P.G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory & Cognition,* **23**, 681-696.

Simoncelli, E.P. (1999). Image and Multi-scale Pyramid Tools [Computer software]. NewYork: Author.

Simons, D., & Wang, R.F. (1998). Perceiving real-world viewpoint changes. *Psychological Science.* **9**, 315-320.

Tanaka, J., & Gauthier, I. (1997). Expertise in object and face recognition. In R.L. Goldstone, D.L. Medin, & P.G. Schyns (Eds.), *Perceptual Learning*. San Diego: Academic Press.

Tanaka, J.W., & Presnell, L.M. (1999). Color diagnosticity in object recognition. *Perception & Psychophysics*, **61**, 1140-1153.

Tanaka, J., & Taylor, M.E. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology,* **15**, 121-149.

Tarr, M.J., & Bülthoff, H.H. (1995). Is human object recognition better described by geon structural descriptions or by multiple views? *Journal of Experimental Psychology: Human Perception and Performance*, **21**, 1494-1505.

Tarr, M.J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, **21**, 233-282.

Tarr, M.J., & Pinker, S. (1991). Orientation-dependent mechanisms in shape recognition: Further issues. *Psychological Science*, **2**, 207-209.

Troje, N. & Bülthoff, H.H. (1996) Face recognition under varying pose: The role of texture and shape. *Vision Research*, **36**, 1761-1771.

Ullman, S. (1998). Three-dimensional object recognition based on the combination of views. *Cognition*, **67**, 21-44.

Yarbus, A.L. (1965). *Role of eye movements in the visual process*. Nauka: Moscow, USSR.

Taxonomic relations and cognitive economy in conceptual organization

Anna M. Borghi (borghi@psibo.unibo.it)
Department of Psychology, 5 Viale Berti Pichat
Bologna, 40127 ITALY

Nicoletta Caramelli (ncaramelli@psibo.unibo.it)
Department of Psychology, 5 Viale Berti Pichat
Bologna, 40127 ITALY

Abstract

This study, performed on children aged 5, 8, and 10, and on adults, deals with conceptual knowledge organization using a word association task. Participants were presented with concept nouns at superordinate, basic, and subordinate level. Their productions were coded according to 4 kinds of relations: taxonomic, thematic, attributive, and evaluative relations. The following results were found at all the considered ages: a. not only lower but also superordinate level concepts elicit attributive relations; b. the production of thematic relations outnumbers that of taxonomic relations thus showing that there is no thematic to taxonomic shift. These results suggest a revision of the criteria on which cognitive economy rests in knowledge organization that will probably lead to a more complex view of the cognitive economy principle.

Introduction

Many models of conceptual organization, from the classical theory and its revised version, the binary theory, to the prototype and some connectionist models, but not the exemplar models, rest on the assumption that the cognitive economy principle underlies both the storing and the retrieval of conceptual information. Accordingly, concepts are defined by the properties and the attributes that establish their identity as well as by their relationships. It is the hierarchical organization of taxonomic relations binding them together that allows people to infer the shared properties and attributes which make the conceptual network coherent.

In this perspective, cognitive development is a progression towards the attainment of this taxonomically and hierarchically organized knowledge structure. During development children undergo a thematic - to - taxonomic shift that is responsible for their mastering their dealings with the environment. This is possible thanks to a well structured knowledge organization that rests on the hierarchical array of taxonomic relations. Thus, cognitive development involves the transition from a contextual or thematic knowledge, based on the acquisition of recurrent properties of objects and events directly experienced, to a more abstract knowledge based on the taxonomic relations responsible for the way objects and events are grouped into categories (Lucariello & Nelson, 1985; Lucariello, Kyratzis & Nelson, 1992).

It is reasonable to argue, however, that also information that is not necessarily inferred from the hierarchical levels of concept plays a relevant role in conceptual organization. It has been claimed that conceptual knowledge is situated and contextually set up (Tschacher & Scheier, 1999). In everyday life, at different times, people perceive different objects in the same spatial context. On the other hand, the same object can be perceived in different spatial contexts (Barsalou, 1993, 1999; Barsalou & Hale, 1993; Barsalou & Prinz, 1997).

In cognitive development these arguments lead to the questioning of the traditional cognitive economy principle based on the hierarchical organization of conceptual knowledge. Some authors have pointed out that the kind of task given to children biases their preference for thematic or taxonomic relations (Waxman & Kosowsky, 1990; Waxman & Namy, 1997) and that thematic relations still play a role in both older children's and adults' conceptual organization (Markman, 1989; Sell, 1992). Therefore there should be no reason to suppose that children undergo a thematic - to - taxonomic shift, i.e. that, with age, the taxonomic organization of conceptual knowledge replaces thematic knowledge (Osborne & Calhoun, 1998).

Moreover, other approaches assume that even abstract information is grounded in perception and action (Barsalou, 1999; Glenberg, 1997; Mandler, 1992; 1997; Smith, 1995; Smith & Heise, 1992). However, these views based on the role of perception and action in shaping conceptual knowledge deal with difficulty with superordinate level concepts, i.e. concepts, as 'animal', that do not refer to a particular, concrete referent.

This study is aimed at shedding some light on two points: a. does thematic knowledge concur with the taxonomic organization of concepts in shaping knowledge in children as well in adults, instead of

loosing its relevance? b. Can superordinate level concepts, not referring to concrete objects, convey perceptual information?

In order to answer these questions, the following hypotheses can be advanced:

1. Hierarchical Levels: if perceptual and action views of categorization hold, superordinate level concepts, not referring to concrete objects, should convey not only abstract but also perceptually and action grounded information. As their activation should involve the activation of their exemplars, they are also expected to elicit attributive relations (Callanan, Repp, McCarthy, & Latzke, 1994; Markman, 1985). Moreover, the relations elicited by basic and subordinate level concepts should be more similar to each other than those elicited by superordinate level concepts as both of them refer to concrete objects. Therefore their activation should elicit mainly attributive and thematic relations.

2. Conceptual Relations: assuming the perceptual and action views of categorization, the perceptually and contextually grounded thematic and attributive relations should characterize not only children's but also adults' conceptual organization. Therefore no thematic - to - taxonomic shift should occur.

These hypotheses were tested on children aged 5, 8, 10 in experiment 1, and on adults in experiment 2. Participants were given a word association task to be performed on concepts at different hierarchical levels, i.e. superordinate (e.g. animal), basic (e.g. dog), and subordinate (e.g. hunting dog) level. The conceptual relations produced by participants were classified as taxonomic, thematic, attributive, and evaluative (Chaffin, 1992; 1997).

Experiment 1

Method

Participants One-hundred and twenty middle class children, 40 aged 5, 40 aged 8 and 40 aged 10, living in Bologna and the surrounding area took part in the study.

Materials To maintain children's attention, only 9 concept-nouns were selected, 3 superordinate (e.g. furniture), 3 basic (e.g. chair), and 3 subordinate level concepts (e.g. high chair). The basic level was defined by the common shape criterion according to which basic level concepts whose members share a common shape are the most inclusive ones (Lassaline, Wisniewski, & Medin, 1992). The superordinate and the subordinate levels were defined respectively as more general and inclusive and more specific than the basic level. All the selected concepts were countable nouns.

Procedure The children were interviewed, one at a time, in their kindergarten or school. They were presented with a booklet. On each page there was a circle and at its center there was a concept-noun. They were asked to say and then to write on the circle from a minimum of 5 to a maximum of 10 associations to each concept-noun. The circle was supposed to prevent children from producing chain-like associations. The free association task, already used with success also with very young children (Nelson, 1986; Sell, 1992), was introduced to the children as if it were a game. At the end of the task, to better assess the intended meaning of the produced associations, the experimenter asked the children why they had said or written what they had said or written for each produced association and tape-recorded their answers.

Coding The data were transcribed and coded by two independent judges (2% cases of disagreement solved after brief discussion), according to 4 different kinds of relations (A) Taxonomic Relations: including superordinate, subordinate, and co-ordinate relations: e.g. bird-animal', bird-parrot', 'sparrow-parrot'. The production of taxonomic relations does not imply that children master class inclusion (Greene, 1994); (B) Thematic Relations: including spatial (physician-hospital), temporal (bird-spring), action (agent, object, and action) (bird-flies), function (chair-to sit on), and event relation when the child reported a whole story; (C) Attributive Relations: including partonomic (chair-leg), property (chair-brown), and matter relation (chair-wood); (D) Evaluative Relations: including meta-linguistic evaluations (oculist-I don't like him/her) as well as stereotyped associations (bird-airplane).

The relations that could not be included in the previous categories, 2% of the relations produced, were not analyzed.

Data analysis and results

To test the hypothesis 1, the percentage of the 4 kinds of the produced relations was computed for each age level and for each hierarchical level (see Table 1, 2, and 3). Three Correspondence Analyses were performed in order to verify whether, at each age level, the distribution of the frequencies of the 4 groups of relations varied across the hierarchical levels. In this analysis, based on the chi square test, the frequencies of the produced relations, from which a broad data matrix is derived, allow the identification of the weight of the different coded dimensions and their graphical representation. On the graph, the geometrical proximity of the points shows the degree of their association and the similarity of their distribution (Hair, Anderson, Tatham & Black, 1992; Greenacre & Blasius, 1994).

The analysis on the relations produced by 5 year olds shows that on the first dimension (explaining 85% of the total variance) superordinate level concepts, characterized by taxonomic relations, differ from subordinate relations characterized by thematic relations. On the second dimension (explaining 15% of the variance), attributive relations, that characterize basic level concepts, differ from evaluative relations.

Table 1: Five year olds. Percentage of the 4 kinds of relation at each hierarchical level.

Relation	Sup	Bas	Sub
Taxonomic	33	20	11
Thematic	48	61	69
Attributive	10	15	12
Evaluative	9	4	9

The analysis on the relations produced by 8 year olds shows that on the first dimension, which explains 97% of the total variance, superordinate level concepts, characterized by taxonomic relations, differ from both basic and subordinate level concepts characterized by thematic relations.

Table 2: Eight year olds. Percentage of the 4 kinds of relation at each hierarchical level.

Relation	Sup	Bas	Sub
Taxonomic	33	13	11
Thematic	34	53	49
Attributive	29	32	36
Evaluative	4	3	4

The analysis on the relations produced by 10 year olds shows that on the first dimension, which explains almost all the variance (99%), superordinate level concepts, characterized by taxonomic relations, differ from subordinate level concepts characterized by thematic relations.

Table 3: Ten year olds. Percentage of the 4 kinds of relation at each hierarchical level.

Relation	Sup	Bas	Sub
Taxonomic	30	15	12
Thematic	26	42	46
Attributive	32	32	32
Evaluative	12	11	9

As the percentages and the correspondence analyses suggest, at all the ages considered, the main difference between superordinate and lower level concepts does not depend on the production of attributive and evaluative relations, but on the production of taxonomic and thematic relations. Superordinate level concepts elicit more taxonomic and less thematic relations than the lower level concepts. However, it is worth noting that superordinate level concepts elicit as many attributive relations as the other hierarchical levels. This could mean that perceptual information that is involved in attributive relations is conveyed not only by lower but also by superordinate level concepts. This result brings new evidence to the perceptual and action views of conceptual knowledge organization. Superordinate level concepts elicit mainly taxonomic relations at the subordinate level, i. e. instantiations (Heit & Barsalou, 1996), (98%, 99%, and 97% in 5, 8, and 10 year olds respectively), thus showing their 'plural force' (Markman, 1985; 1989; Murphy & Wisniewski, 1989). The same is found also in basic level concepts (88%, 72%, and 76% respectively), though the percentage of instantiations decreases consistently. Subordinate level concepts, instead, elicit mostly items at the superordinate level (55%, 53%, and 52% respectively).

To test hypothesis 2 a Correspondence Analysis was performed on the 4 kinds of relation crossed with the age levels. The first dimension, explaining 76% of the variance, shows the difference between 5 year olds, who produce mostly thematic relations, and 10 year olds who produce mostly attributive relations. The far less relevant second dimension, explaining 24% of the variance, shows the difference between 8 and 10 year olds as these last produce evaluative relations (see Figure 1).

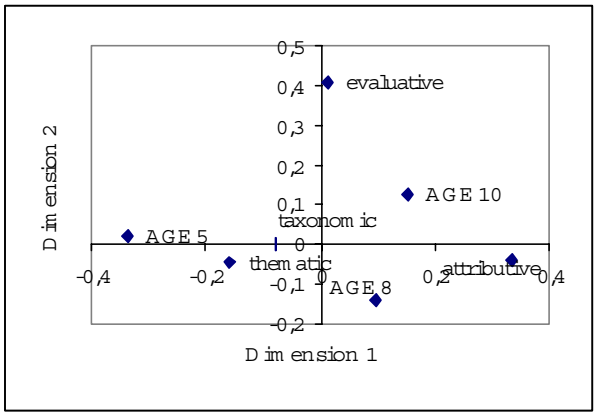


Figure 1: Correspondence Analysis on the 4 kinds of relation at each age level. Dimension 1 = 76% - Dimension 2 = 24%.

Even though the production of thematic relations decreases with age, at all the age levels the production of thematic relations outnumbers that of the other relations, while that of taxonomic relations does not consistently change (see Table 4). Thus, there seems to be no evidence of a thematic - to - taxonomic shift.

Furthermore, the production of perceptually grounded relations, i.e. attributive relations, increases consistently from 5 to 8 years.

Table 4: Percentage of the 4 kinds of relation produced at each age level.

Relation	Age 5	Age 8	Age 10
Taxonomic	23	20	19
Thematic	57	45	38
Attributive	12	32	32
Evaluative	8	4	10

This last result can be interpreted in the light of the growing analytical attitude children develop with schooling. The production of evaluative relations drops in 8 year olds and then increases again in 10 year olds.

Experiment 2

Experiment 2 was a replica of Experiment 1 but with adult participants and hence with slight variations. The aim was to test whether the trends found in Experiment 1 in 10 year old children are maintained in adults.

Method

Participants Two hundred students of the University of Bologna volunteered for their participation.

Materials Twenty-seven concepts were selected; each presented at the 3 hierarchical levels.

Procedure The procedure was the same as in Experiment 1 with minor adaptations for the adult participants. The task was not presented as a game and participants were allowed to write associations of associations. When presented with 'flower', for example, they could think of 'geranium' and had to write this word linked to 'flower'. If 'geranium' made them think of 'vase', they had to write 'vase' linked to 'geranium'. Thus, it was possible to distinguish between direct and indirect associations elicited by the given concept nouns. Participants could produce as many associations as they wanted in a maximum time of 5 minutes for each concept. Only direct associations were analyzed in this research

Coding The coding procedure and codes were the same as those used in Experiment 1. The two judges solved 8% cases of disagreement after brief discussion.

Data analysis and results

In order to test hypothesis 1, the percentages of the 4 kinds of the produced relations were computed for each hierarchical level of the concept-nouns (see Table 5). A Correspondence Analysis was performed on the frequencies of the relations produced crossed with the hierarchical levels of the concept-nouns. On the first dimension, explaining almost all the variance (98%), superordinate level concepts, that elicit taxonomic relations, differ from subordinate level concepts that elicit thematic relations.

Table 5: Adults. Percentage of the 4 kinds of relation at each hierarchical level.

Relation	Sup	Bas	Sub
Taxonomic	30	19	15
Thematic	29	37	42
Attributive	21	21	23
Evaluative	20	22	21

Both percentages and Correspondence Analysis replicate the same pattern of results as that found in children. Superordinate level concepts elicit more taxonomic relations and less thematic relations than lower level concepts. Again, attributive relations are elicited in the same proportion by superordinate level concepts as by lower level concepts.

Comparison between Experiment 1 and 2

In order to test hypothesis 2, the relations produced by children at the different age levels and by adults were compared.

As Figure 2 shows, the production of thematic relations gradually decreases between 5 and 10 years, and then it increases again. It is also worth noting that thematic relations are the most frequently produced by both children and adults. The production of taxonomic relations, instead, is more stable across the age levels than that of thematic relation. A greater variability is found in the production of attributive relations.

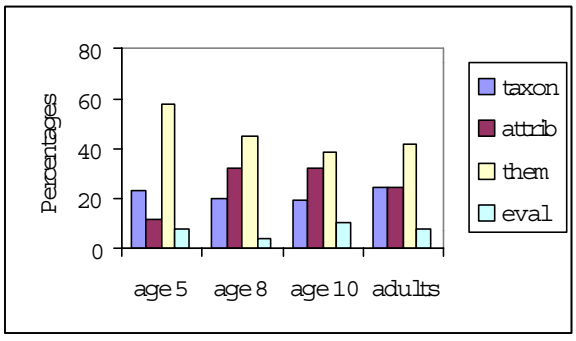


Figure 2: Percentage of the 4 kinds of relations at each age level.

A Correspondence Analysis was performed crossing the 4 groups of relations produced with the 4

different age levels (children aged 5, 8, and 10, and adults). The first dimension (explaining 67% of the total variance) shows the difference between 5 year olds, who produce thematic relations, and 10 year olds who produce attributive relations. On the second dimension (explaining 24% of the variance) both 5 and 8 year olds who produce attributive and thematic relations, differ from 10 year olds who produce taxonomic relations (see Figure 3). In this analysis adult's productions have no weight; this means that the pattern of relations they produce does not differ from that produced by children.

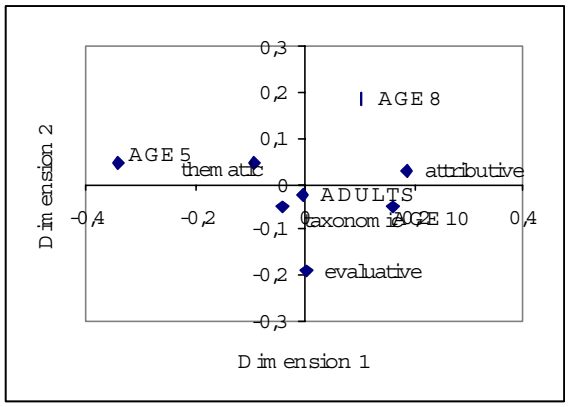


Figure 3: Correspondence Analysis on the 4 kinds of relations at each age level. Dimension 1 = 67% ; Dimension 2 = 24%

General Discussion

The results verify the two hypotheses set forth at the beginning and support perceptual and action-based views of conceptual organization. They show that:

1. At each considered age, superordinate, basic and subordinate level concepts convey attributive, i. e. perceptually grounded, information. Such finding is consistent with the view advanced by Smith & Heise (1992) and Jones & Smith (1998) that perceptual properties are relevant in distinguishing between different kinds of superordinate level concepts. For example, both textural information and the difference between biological and non-biological movement help to distinguish between very general superordinate concepts such as artifacts and natural kind concepts.

The fact that superordinate level concepts convey more taxonomic and less thematic and contextual information than lower level concepts can be a consequence of the instantiation principle (Heit & Barsalou, 1996). The activation of superordinate level concepts elicit information about their exemplars thanks to their eliciting both attributive relations and instantiations of the subordinate kind. This result is consistent with the widely acknowledged result that superordinate level concepts activate their exemplars (Callanan & al., 1994; Markman, 1985; 1989; Murphy & Wisniewski, 1989).

2. Contextual and thematic information plays a relevant part in organizing knowledge not only in children but also in adults. There is no evidence of a shift from taxonomic - to - thematic knowledge organization at least using a production task, as is the case in this study. In fact, the production of taxonomic relations does not change across the age levels. At all the ages considered, concepts convey more thematic than taxonomic information while the weight of attributive information becomes more relevant with age in shaping conceptual knowledge. This finding is supported also by recent evidence on the lack of a consistent preference for either thematic or taxonomic relations by pre-school children (Osborne & Calhoun, 1998; Waxman & Namy, 1997) and on conceptual flexibility and variability (Smith & Samuelson, 1997).

This research shows that the cognitive economy principle, resting on a hierarchical organization of taxonomic relations, is not able to handle the way conceptual knowledge is really organized (for a similar conclusion see Sloman, 1998). The cognitive economy principle has to be revised so that it can handle all the nuances knowledge inherits from our complex dealings with the environment. This does not mean that abstraction has no part in knowledge organization, it only means that even abstract knowledge originates from direct experience.

Acknowledgments

We thank Larry Barsalou for useful comments on some aspects of this research and Angela Montanari for statistical advice. The research reported in this paper was supported by grants from CNR n. 9504047, 9601807, 9700385, 9800572CT11 to the second author.

References

Barsalou, L. W. (1993). Flexibility, structure, and linguistic vagary in concepts: Manifestations of a compositional system of perceptual symbols. In A. F. Collins, S. E. Gathercole, M. A. Conway, P. E. Morris (Eds.), Theories of memory. Hillsdale, NJ: Lawrence Erlbaum Associates.

Barsalou, L. W. (1999). Perceptual Symbol Systems. Behavioral and Brain Sciences, 22, 577-609.

Barsalou, L. W., & Hale, C. R. (1993). Components of conceptual representations: from feature list to recursive frames. In I. Van Mechelen, R. S. Michalski (Eds.). Categories and concepts: Theoretical views and inductive data analysis. London: Academic Press.

Barsalou, L. W., & Prinz, J. J. (1997). Mundane creativity in perceptual symbol systems. In T. B. Ward, S. M. Smith & J. Vaids (Eds.). Creative thought: An investigation of conceptual structures

and processes. Washington, DC: American Psychological Association Press.
Callanan, M.A., Repp, A.M., McCarthy, M.G., & Latzke, M.A. (1994). Children's hypotheses about word meaning: Is there a basic level constraint? Journal of Experimental Child Psychology, 57, 108-138.
Chaffin, R. (1992). The concept of a semantic relation. In E. Kittay & A. Lehrer (Eds.), Frames, Fields and Contrasts: New Essays in Lexical and Semantic Organization. Hillsdale, NJ: Lawrence Erlbaum Associates.
Chaffin, R. (1997). Associations to unfamiliar words: Learning the meanings of new words. Memory and Cognition, 25, 2, 203-226.
Glenberg, A.M. (1997). What memory is for. Behavioral and Brain Sciences, 20, 1-55.
Greene, T.R. (1994). What Kindergartners Know about Class Inclusion Hierarchies. Journal of Experimental Child Psychology, 57, 72-88.
Greenacre, M., & Blasius, J. (Eds) (1994). Correspondence analysis in the social sciences: recent developments and applications. London: Academic Press.
Hair, J.F., Anderson, R.E., Tatham, R.L., & Black, W.C. (1992). Multivariate data analysis. New York: MacMillan.
Heit, E., & Barsalou, L.W. (1996). The instantiation principle in natural categories. Memory, 4, 4, 413-451.
Jones, S.S. & Smith, L.B. (1998). How children name objects with shoes. Cognitive Development, 13, 323-334.
Lassaline, M.E., Wisniewski, E.J., & Medin, D.L. (1992). Basic levels in artificial and natural categories. In B. Burns (Ed.), Percepts, concepts and categories. Amsterdam: Elsevier.
Lucariello, J., Kyratzis, A., & Nelson, K. (1992). Taxonomic knowledge: what kind and when? Child Development, 63, 978-998.
Lucariello, J., & Nelson, K. (1985). Slot-filler categories as memory organizers for young children. Developmental Psychology, 21, 272-282.
Mandler, J.M. (1992). How to build a baby: II. Conceptual primitives. Psychological Review, 99, 587-604.
Mandler, J.M. (1997). Development of categorization: Perceptual and conceptual categories. In G. Bremner, A. Slater, & G. Butterworth (Eds.), Infant Development. Recent Advances. Hove, UK: Psychology Press.
Markman, E.M. (1985). Why superordinate category terms can be mass nouns. Cognition, 19, 311-353.
Markman, E.M. (1989). Categorization and naming in children: Problems of induction. Cambridge, MA: MIT Press.
Murphy, G.L., & Lassaline, M.E. (1997). Hierarchical Structure in Concepts and the Basic Level of Categorization. In K. Lamberts & D. Shanks (Eds.), Knowledge, Concepts, and Categories. Hove, UK: Psychology Press.
Murphy, G.L., & Wisniewski, E.J. (1989). Categorizing objects in isolation and in scenes: What a superordinate is good for. Journal of Experimental Psychology: Learning, Memory and Cognition, 15, 572-586.
Nelson, K. (1986). Event knowledge. Structure and function in development. Hillsdale, NJ: Erlbaum.
Obsborne, G.J., & Calhoun, D.O. (1998). Themes, Taxons, and Trial Types in Children's Matching to Sample: Methodological Considerations. Journal of Experimental Child Psychology, 68, 35-50.
Sell, M.A. (1992). The development of children's knowledge structure: events, slots, and taxonomies. Journal of Child Language, 19, 659-676.
Sloman, S.A. (1998). Categorical Inference Is Not a Tree: The Myth of Inheritance Hierarchies. Cognitive Psychology, 35, 1, 1-33.
Smith, L.B. (1995). Stability and Variability: the geometry of children's novel-word interpretations. In F.D. Abraham & A.R. Gilgen (Eds.), Chaos theory in psychology. London, UK: Praeger.
Smith, L.B., & Heise, D. (1992). Perceptual similarity and conceptual structure. In B. Burns (Ed.), Percepts, concepts and categories. Amsterdam: Elsevier.
Smith, L.B. & Samuelson, L.L. (1997). Perceiving and Remembering: Category Stability, Variability and Development. In K. Lamberts & D. Shanks (Eds), Knowledge, Concepts, and Categories. Hove, UK: Psychology Press.
Tschacher, W., & Scheier, C. (1999). Situated and self-organizing cognition. In P. Van Loocke (Ed.), The nature of concepts. London, UK: Routledge.
Waxman, S.R., & Kosowsky, T.D. (1990). Nouns mark category relations: Toddlers' and preschoolers' word-learning biases. Child Development, 61, 1461-1473.
Waxman, S.R., & Namy, L.L. (1997). Challenging the notion of a thematic preference in children. Developmental Psychology, 33, 555-567.
Wisniewski, E.J., Imai, M., & Casey, L. (1996). On the equivalence of superordinate concepts. Cognition, 60, 269-298.

The Roles of Body and Mind in Abstract Thought

Lera Boroditsky (lera@psych.stanford.edu)
Department of Psychology, Jordan Hall, Bldg420
Stanford, CA 94305-2130 USA

Michael Ramscar (michael@dai.ed.ac.uk)
University of Edinburgh, 2 Buccleuch Place
Edinburgh, EH8 9LW, Scotland

Michael C. Frank (mcfrank@stanford.edu)
Department of Psychology, Jordan Hall, Bldg420
Stanford, CA 94305-2130 USA

Abstract

How are we able to think about things we've never seen or touched? We demonstrate that abstract knowledge is built analogically from more experience-based knowledge. People's understanding of the abstract domain of time, for example, is so intimately dependent on the more experience-based domain of space, that when people make a an air journey or bet on a racehorse, they also unwittingly (and dramatically) change their thinking about time. Further, it appears that abstract thinking is built on representations of more experience-based domains that are functionally separable from those involved directly in sensorimotor experience itself.

How are we able to think about things we've never seen or touched? Whether we are theorizing about invisible forces, studying the behaviors of atoms, or trying to characterize the nature of private mental experience, much scientific progress depends on generating new ways of describing and conceptualizing phenomena which are not perceivable through the senses. We face the same problems in everyday life with abstract notions like time, justice, and love. How do we come to represent and reason about abstract domains despite the dearth of sensory information available about them?

One suggestion is that abstract domains are understood through analogical extensions from richer, more experience-based domains (Boroditsky, 2000; Gentner et al., in press; Gibbs, 1994; Holyoak & Thagard, 1995; Lakoff & Johnson, 1980, 1999). This experience-based structuring view can be formulated in several strengths. A very strong "embodied" formulation might be that knowledge of abstract domains is tied directly to the body such that abstract notions are understood directly through image schemas and motor schemas (Lakoff & Johnson, 1999). A milder view might be that abstract knowledge is based on representations of more experience-based domains that are functionally separable from those directly involved in sensorimotor experience.

The studies reported in this paper show that people's understanding of the abstract domain of time is substrated by their knowledge and experiences in the more concrete domain of space[1]. In fact, people's representations of time are so intimately dependent on space that when people engage in particular types of spatial thinking (e.g., embarking on a train journey, or urging on a horse in a race), they unwittingly also change how they think about time. Further (and contrary to the very strong embodied view), it appears that abstract thinking is built on *representations* of more experience-based domains, and not necessarily on the physical experience itself.

Suppose you are told that next Wednesday's meeting has been moved forward two days. What day is the meeting now that it has been rescheduled? The answer to this question depends on how you choose to think about time. If you think of yourself as moving forward through time (the ego-moving perspective), then moving a meeting "forward" is moving it further in your direction of motion—that is, from Wednesday to Friday. If, on the other hand, you think of time as coming toward you (the time-moving perspective), then moving a meeting "forward" is moving it closer to you—that is, from Wednesday to Monday (Boroditsky, 2000; McGlone & Harding, 1998; McTaggart, 1908). In a neutral context, people are equally likely to think of themselves as moving through time as they are to think of time as coming toward them, and so are equally

[1] This paper only examines one aspect of how people think about time. The domain of time comprises an incredibly complex, heterogeneous and sophisticated set of representations, and much more investigation will be necessary to properly characterize the domain as a whole.

likely to say that the meeting has been moved to Friday (the ego-moving answer) as to Monday (the time-moving answer) (Boroditsky, 2000; McGlone & Harding, 1998).

But where do these representations of time come from? Is thinking about moving through time based on our more concrete experiences of moving through space? If so – if representations of time are indeed tied to representations of space – then getting people to think about space in a particular way should also influence how they think about time.

Study 1

To investigate the relationship between spatial thinking and people's thinking about time, we asked 239 Stanford undergraduates to fill out a one-page questionnaire that contained a spatial prime followed by the ambiguous "Next Wednesday's meeting..." question described above. The spatial primes (shown in Figure 1) were designed to get people to think about themselves moving through space on an office-chair (see Figure 1a), or of making an office-chair come toward them through space (see Figure 1b). In both cases, participants were asked to imagine how they would "manuever the chair to the X," and to "draw an arrow indicating the path of motion." The left-right orientation of the diagrams was counterbalanced across subjects. Immediately after subjects completed the spatial prime, they were asked the ambiguous "Next Wednesday's meeting has been moved forward two days" question described above. We were interested in whether subjects would think differently about time right after imagining themselves as moving through space, or imagining things coming toward them.

Figure 1a: Spatial prime used in Study 1. Participants were given the following instructions "Imagine you are the person in the picture. Notice there is a chair on wheels, and a track. You are sitting in the chair. While sitting in the chair, imagine how you would manuever the chair to the **X**. Draw an arrow indicating the path of motion."

Figure 1b: Spatial prime used in Study 1. Participants were given the following instructions "Imagine you are the person in the picture. Notice there is a chair on wheels, and a track. You are holding a rope attached to the chair. With the rope, imagine how you would manuever the chair to the **X**. Draw an arrow indicating the path of motion."

As predicted, people used primed spatial information to think about time. Subjects primed to think of objects coming toward them through space, were much more likely to think of time as coming toward them (67% said Wednesday's meeting had moved to Monday), than they were to think of themselves as moving through time (only 33% said the meeting has moved to Friday). Subjects primed to think of themselves as moving through space showed the opposite pattern (only 43% said Monday, and 57% said Friday), chi-square=13.3, $p<.001$. It appears that people's thinking about time is indeed tied to their spatial thinking. This raises a further question: do people unwittingly change their thinking about time during everyday spatial experiences and activities (not just when processing specially designed spatial primes in a laboratory setting)?

Study 2: The Lunch-line

To investigate the relationship between spatial experience and people's thinking about time, we asked 70 people waiting in a lunch-line the ambiguous question about Wednesday's meeting described above. The lunch-line was for a café in the basement of Stanford's Psychology department. The line is usually about 50 meters long, but moves quickly with a waiting time of about 10 minutes. After participants answered our ambiguous question, we asked them how long they felt they had waited in line, and also recorded which quartile of the line they were in when interviewed. This second index served as an objective measure of how much forward motion in line people had experienced (with people furthest along in line having experienced the most motion). We were interested in whether the spatial experience of moving forward in a line would make people more likely to also think of themselves as moving forward in time (as opposed to thinking of time as coming toward them).

As predicted, the further along in line people were (the more forward spatial motion they had experienced), the more likely they were to also think of themselves as moving through time (to say the meeting had been moved to Friday), $r=.33$, $p<.005$ (see Figure 2). People's estimates of their waiting time were also predictive of their answers to the question about next Wednesday's meeting ($r=.26$, $p<.05$). Interestingly,

people's estimates of their waiting time were less predictive of their answer to the "Next Wednesday's meeting…" question than was their spatial position in line. When the effect of spatial position was controlled for, people's estimates of their waiting time were no longer predictive of their answers to the ambiguous question about time, r=.05, p=.67. It appears that spatial position in line (and hence the amount of forward spatial motion that a person had just experienced) was the best predictor of their thinking about time.

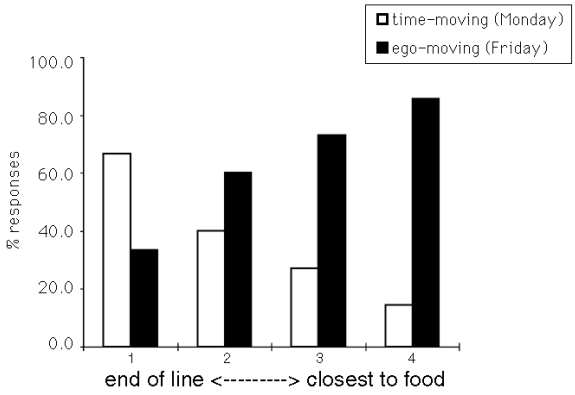


Figure 2: Responses of 70 people waiting in a lunch-line. Responses are plotted by position in line (from the end quartile of the line to the quartile closest to the food). The further along in line people were (and hence the more forward spatial motion they had experienced), the more likely they were to take the ego-moving perspective on time (say that next Wednesday's meeting has been "moved forward" to Friday).

In the next study we were interested in whether spatial motion per se was necessary, or whether simply thinking about or anticipating a journey would be enough to influence how people think about time.

Study 3: The Airport

To investigate the relationship between spatial experience and people's thinking about time, we asked 333 visitors to San Francisco International Airport the ambiguous question about Wednesday's meeting described above. After the participants answered, we asked them whether they were waiting for someone to arrive, waiting to depart, or had just flown in. We were interested in two things: (1) whether a lengthy experience of moving through space would make people more likely to take the ego-moving perspective on time (think of themselves as moving through time as opposed to thinking of time as coming toward them), and (2) whether the actual experience of motion was necessary to change one's thinking about time, or if just thinking about motion was enough.

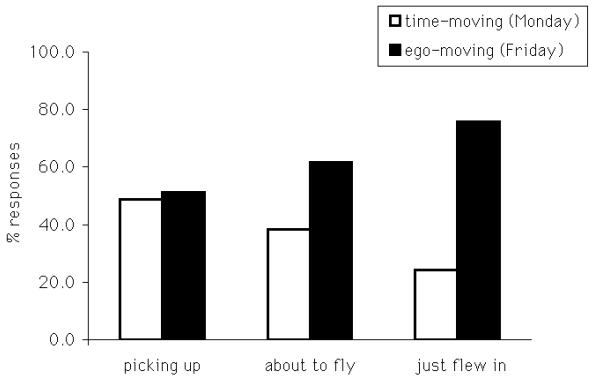


Figure 3: Responses of 333 people queried at the airport. People who had just flown in were most likely to produce an ego-moving response (say that next Wednesday's meeting has been "moved forward" to Friday).

As shown in Figure 3, people who had just flown in were much more likely to take the ego-moving perspective (think of themselves as moving through time and answer Friday) (76%) than people who were just waiting for someone to arrive (51%), χ^2=14.3, p<.01. Further, even people who hadn't yet flown, but were only waiting to depart were already more likely to think of themselves as moving through time (62%), χ^2=4.3, p<.05 (when compared to people waiting for someone to arrive). This set of findings suggests that (1) people's ideas about time are indeed intimately related to their representations of space, and (2) just thinking about spatial motion is sufficient to change one's thinking about time. But this also raises an interesting question: why were people who had just flown in more likely to take an ego-moving perspective than people who were only about to depart? Was it because they had spent more time actually moving through space, or was it just because they had had more time to think about it?

Study 4: The Train

To investigate this further, we posed the ambiguous question about Wednesday's meeting to 219 patrons of CalTrain (a train-line connecting San Francisco and San Jose). Of these, 101 were people waiting for the train, and 118 were passengers actually on the train. All of them were seated at the time that they were approached by the experimenter. After participants answered our question, we asked them about how long they had been waiting for (or been on) the train, and how much further they had to go. Participants indicated their

waiting/travel times using a multiple-choice questionnaire.

First, we found that both people waiting for the train and people actually riding on the train were more likely to take the ego-moving perspective (63%) than the time-moving perspective (37%), χ^2=13.9, p<.01. Interestingly, the data from people waiting for the train didn't look any different from those of people actually on the train (61% and 64% ego-moving response respectively), suggesting that it is not the experience of spatial motion per se, but rather thinking about spatial motion that underlies our representation of time.

We then examined people's responses based on how long they had been waiting for the train (see Figure 4). The longer people sat around thinking about their journey, the more likely they were to take the ego-moving perspective for time. People who had waited less than a minute were equally likely to think of themselves as moving through time as they were to think of time as coming toward them. People who had had five minutes of anticipating their journey, were much more likely to take the ego-moving perspective on time (68%), χ^2=4.5, p<.05 (when compared to people waiting less than a minute).

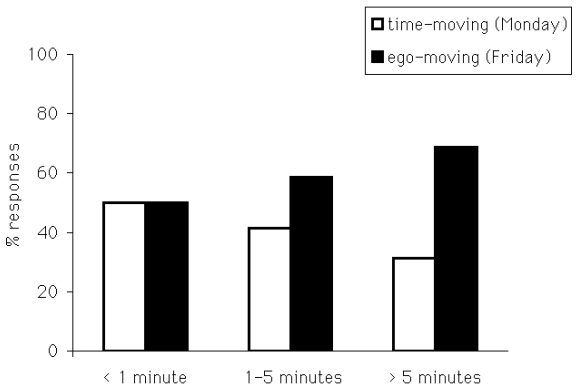


Figure 4: Responses of 101 people waiting for the train plotted by time spent waiting. The more time people had to anticipate their journey, the more likely they became to adopt the ego-moving perspective on time (say that next Wednesday's meeting has been "moved forward" to Friday).

Finally, we analyzed the responses of people on the train based on whether they had answered our ambiguous time question at the beginning, middle, or at end of their journey. We conjectured that people should be most involved in thinking about their journey when they had just boarded the train, or when they were getting close to their destination. In the middle of their journey, people tend to relax, read, talk loudly on cell-phones, and otherwise mentally disengage from being on the train.

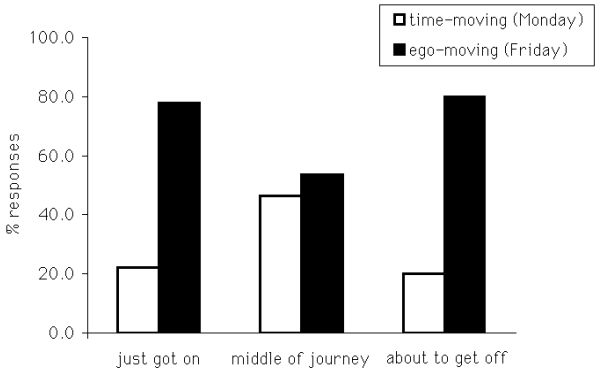


Figure 5: Responses of 118 passengers on the train plotted by point in journey. People became much more likely to adopt the ego-moving perspective for time (say that next Wednesday's meeting has been "moved forward" to Friday) when they were most engaged in thinking about their spatial journey (at the beginnings and ends of the trip). In the middle of their journey, people were about equally likely to adopt the ego-moving perspective (say the meeting has been "moved forward" to Friday) as the time-moving perspective (say the meeting has been "moved forward" to Monday).

Amazingly, people's biases for thinking about time mimicked this pattern of engaging and disengaging from spatial thinking perfectly (see Figure 5). Within five minutes of getting on the train, people were very likely to take the ego-moving perspective on time (78%), χ^2=6.38, p<.02 (when compared to people in the middle of their journey). People were also very likely to take the ego-moving perspective when they were within ten minutes of arriving at their destination (80%), χ^2=5.63, p<.02 (when compared to people in the middle of their journey). Passengers in the middle of their journey, however, showed no ego-moving bias. They were just as likely to think of themselves as moving through time (54%), as they were to think of time as coming toward them (46%). Once again it appears that people's thinking about time is substrated by thinking about spatial motion and not necessarily by the experience of motion itself. Although all three groups of passengers were having the same physical experience (all were simply sitting on the train), the two groups that were most likely to be involved in thinking about their journey showed the most change in their thinking about time.

Study 5: The Race-Track

So far, we have only looked at people who themselves were moving or planning to move. Could thinking

about spatial motion have a similar effect even when people are not planning any of their own motion? To investigate this question, we asked the "Next Wednesday's meeting…" question of 53 visitors to the Bay Meadows racetrack. We predicted that the more involved people were in the forward motion of the racehorses, the more likely they would also be to take the ego-moving perspective on time (and say that the meeting has been moved to Friday). After asking people the question about next Wednesday's meeting, we also asked them how many races they had watched that day and how many races they had bet on. Both indices turned out to be good predictors of people's answers to the "Next Wednesday's meeting…" question. As shown in Figure 6, people who hadn't bet on any races were as likely to think of themselves as moving through time (50% said Friday), as they were to think of time as coming toward them (50% said Monday). In contrast, people who had bet on 3 races or more were three times more likely to think of themselves as moving through time (76% said Friday) than they were to think of time as coming toward them (24% said Monday), %), χ^2=12.39, p<.01 (when compared to people who hadn't bet on any races). It appears that simply thinking about forward motion (willing a horse towards a finish line, as opposed to actually planning to go somewhere yourself) is enough to change people's thinking about time.

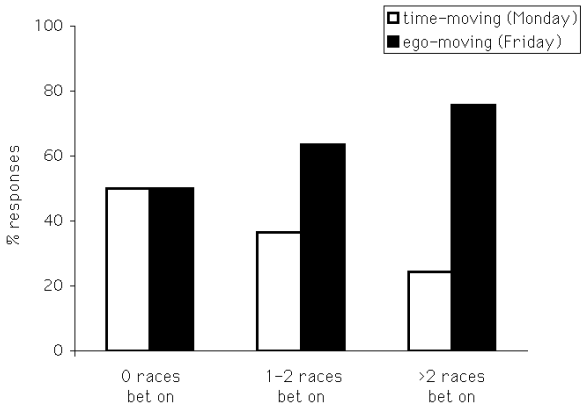


Figure 6: Responses of 53 visitors to the racetrack plotted by number of races bet on. People who had bet on more races (and so were more involved in the forward motions of the racehorses) also became much more likely to adopt the ego-moving perspective for time (say that next Wednesday's meeting has been "moved forward" to Friday).

Study 6: The Office-Chair Rodeo

Thus far we have shown that people's thinking about spatial motion is a good predictor of their thinking about time and that actual spatial motion is not necessary. A further question is whether actual motion is *sufficient* to influence people's thinking about time even in the absence of involved spatial thinking.

To try to address this question, we designed a real-motion version of Study 1 (see Figure 1). We set up a 25 ft track outside of the Stanford University Bookstore and invited students to participate in an "office-chair rodeo." Half of the participants were asked to ride an office chair from one end of the track to the other (the ego-moving prime), and half were asked to rope the chair in from the opposite end of the track (the time-moving prime) (see Figure 1 for a diagram). The track was marked out in the asphalt using colored masking tape, with one end of the track marked in red, and one in yellow. Seventy-eight Stanford undergraduates participated in the study in exchange for lollipops. The verbal instructions were the same in both conditions. Participants riding the chair sat in an office-chair at one end of the track and were asked to "maneuver the chair to the red/yellow line" (whichever was at the opposite end of the track). Participants roping the chair were given a rope that was connected to the office-chair at the opposite end of the track and were likewise instructed to "maneuver the chair to the red/yellow line" (whichever was where the participant was standing).

Immediately after the participant completed the motion task (either riding or roping the chair), they were asked the question about next Wednesday's meeting. Interestingly, performing these spatial motion tasks had no effect on subjects' thinking about time. People riding the chair (actually moving through space), were as likely to think of themselves as moving through time (55% said the meeting would be on Friday), as were people roping the chair (actually making and object move toward them) (58% said the meeting would be on Friday). It appears that just moving through space, is not sufficient to influence people's thinking about time. This finding is especially striking when compared to the findings of Study 1 where just thinking about spatial motion (in the absence of any actual motion) was enough to influence people's thinking about time (see also Boroditsky, 2000).

Discussion

Taken together these studies demonstrate the intimate relationship between abstract thinking and more experience-based forms of knowledge. People's thinking about time is closely linked to their spatial thinking.

When people engage in particular types of spatial thinking (e.g., thinking about their journey on a train, or urging on a horse in a race), they also unwittingly and dramatically change how they think about time. Further, and contrary to the very strong embodied view, it appears that this kind of abstract thinking is built on representations of more experience-based domains that are functionally separable from those directly involved in sensorimotor experience itself.

A further question is how do these relationships between abstract and concrete domains come about in the first place? Surely, some relationships come from correspondences that can be observed in experience. For example, progression in space and time are often correlated -- movements that are longer spatially are also likely to take a longer amount of time. These simple correspondences in experience can then be amplified and built on by language. People often use metaphors from more experienced-based domains to talk about more abstract domains, and often these metaphors go beyond what can be observed in experience. This means that some abstract knowledge might be constructed and shaped by language. In fact, this turns out to be the case. For example, English and Mandarin speakers use different spatial metaphors to talk about time, and this difference in language leads to important differences in the way the two groups think about time (Boroditsky, in press). It follows that to properly characterize abstract thought, it will be important to look not only at what comes from innate wiring and physical experience, but also at the ways in which languages and cultures have allowed us to go beyond these to make us smart and sophisticated as are.

Acknowledgments

The authors would like to thank Amy Jean Reid, Justin Weinstein, and Webb Phillips for their heroic feats of data collection. Correspondence and requests for materials should be addressed to Lera Boroditsky (email: lera@psych.stanford.edu).

References

Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition, 75(1)*, 1-28.

Gentner, D., Bowdle, B., Wolff, P., & Boronat, C. (in press). Metaphor is like analogy. In Gentner, D., Holyoak, K. J., & Kokinov, B. N. (Eds.) (in press). *The analogical mind: Perspectives from cognitive science.* Cambridge, MA: MIT Press.

Gibbs, R.J. (1994). *The poetics of mind: Figurative thought, language, and understanding.* New York, Cambridge University Press.

Holyoak, K.J. & Thagard, P. (1995). *Mental leaps: Analogy in creative thought.* Cambridge, The MIT Press.

Lakoff, G. & Johnson, M. (1980). *Metaphors we live by.* University of Chicago Press, Chicago.

Lakoff, G. & Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought.* Basic Books, New York.

McGlone, M.S. & Harding, J.L. (1998). Back (or forward?) to the future: the role of perspective in temporal language comprehension. *Journal of Experimental Psychology: Learning Memory and Cognition, 24(5)*, 1211-1223.

McTaggart, J. (1908). The Unreality of time. *Mind, 17,* 457-474.

The time-course of morphological, phonological and semantic processes in reading Modern Standard Arabic

Sami Boudelaa (sami.boudelaa@mrc-cbu.cam.ac.uk)

William Marslen-Wilson (marslen.wilson@mrc-cbu.cam.ac.uk)
MRC Cognition and Brain Sciences Unit
15 Chaucer Road, Cambridge CB2 2EF, UK

Abstract

We investigate deverbal noun and verb morphology in Modern Standard Arabic (MSA), using two masked priming experiments in which the morphological, orthographic and semantic relationship between prime and targets are varied in four SOA conditions (32 ms, 48 ms, 64 ms, and 80 ms). Results show that early in the visual processing of MSA deverbal nouns and verbs, the role played by morphological structure (word patterns and roots), is significantly different from that played by orthographic and semantic factors. Additionally, while word pattern effects are transient and overlap with root morpheme effects only in the early stages of processing, effects of the root are reliably present throughout the recognition process.

Introduction

Unlike Indo-European languages where morphemic units are linearly strung one after the other to create new forms, Semitic languages like Modern Standard Arabic (MSA), or Hebrew draw on a non-linear word building principle whereby at least two abstract morphemes are interlaced one within the other (Holes, 1995, Versteegh, 1997). Reflecting this, surface word forms in such languages are traditionally analyzed into word patterns and roots. Word patterns are CV structures, primarily specifying vowels, that provide phonological structure and convey syntactic meaning, while roots consist solely of consonants and convey the broad semantic properties of the surface form (Wright, 1995). For example, the word [katama] "conceal" comprises the word pattern {fa ala}[1] with the syntactic meaning "perfective, active", and the root {ktm} meaning "concealing". Both of these units recur many times in the language in combination with other units. The word pattern {fa ala} is met in other forms like [sakaba] "pour", [daxala] "enter", [faqada] "miss". Likewise, the root {ktm} appears in such forms as [kattama] "cause to hide", [kaatama] "withhold", and [takattama] "keep mum".

Previous research on the use of word patterns and roots during the processing of Hebrew and MSA has yielded interesting, and largely consistent evidence that these units are actively used during processing (Frost, Forster & Deutsch, et al., 1997, Deutsch, Frost & Forster, 1998, Boudelaa & Marslen-Wilson, 2000). For instance, significant word pattern priming effects were found in Hebrew and MSA, although limited to verb morphology in Hebrew, but applying across syntactic class categories in MSA (Deutsch, et al., 1998, Boudelaa & Marslen-Wilson, 2000). Additionally, reliable root priming was also found in Hebrew and MSA nouns and verbs regardless of semantic transparency (Frost et al., 1997, Boudelaa & Marslen-Wilson, 2000). This research into Semitic morphology has provided some of the most compelling evidence in favour of the view that the role played by morphological structure in lexical processing and representation is distinct from form and meaning effects. Furthermore, the effects of word patterns and roots clearly show that bound and indeed disrupted morphemic units do influence processing.

Here we focus on MSA deverbal nouns and verbs and try to go beyond the research reported so far to examine how the prior presentation of a prime word affects lexical decision to a target as a function of (a) the relationship underlying prime and target, and (b) prime display duration. As regards (a), we vary morphological, orthographic and semantic relationships between primes and targets such that the respective contributions of each of these properties can be examined. With respect to (b), we use four display durations (or SOA's), of 32, 48, 64, and 80 ms, to assess the effects of priming across these dimensions of similarity. In an earlier masked priming investigation of Arabic morphology, we found reliable word pattern and root priming effects at an SOA of 48 ms. Since both morphemic units seem to be involved in the processing

[1] This is the conventional notation used to describe word patterns, where the letters "f, , l" are place holders indicating where the first, second and third root letters go when this unit is combined with a word pattern, and the vowels ("a, a" here) indicate which vowels are inserted into the surface CV template.

at such an SOA, we decided to include a shorter SOA of 32 ms, to determine whether word patterns and roots have different processing onsets. We also included the two longer SOA's (64 and 80 ms) to monitor for the life span of the priming likely to be generated by these units. It should be stressed that at SOA's of 32 ms and 48 ms, participants are not aware of the presence of a prime at all, while at 64 and 80 ms, the presence of a prime may be detectable, though never reliably enough to be reported. This means that masked priming performance is relatively insensitive to episodic and strategic confounds. Furthermore, previous research using this paradigm has shown it to be well suited to the study of morphological and orthographic effects at short SOA's (Frost et al., 1997, Forster & Azuma, 2000), and to the investigation of semantic effects at longer SOA's (Perea et al., 1995, Sereno, 1991). Accordingly, apart from minimizing strategic behavior, our choice of a small range of incremental steps in prime durations should allow us to track the dynamics of processing events as they unfold over time, and in a more fine-grained manner than earlier studies using this technique (Rastle et al., 2000, Feldman, 2000). Our hypothesis is that if morphological structure in MSA is playing a role that is genuinely distinct from that played by orthography and semantics, then this should be reflected in the differential priming effects observed in the morphological, orthographic and semantic conditions across the four SOA's. More specifically, word pattern and root effects should be able to emerge earlier than semantic effects, and should be stronger than orthography-driven effects. Additionally, since root morphemes convey semantic meaning, their effects are predicted to be more long-lived than those of the word patterns, which convey syntactic and phonological information. These predictions are tested in Experiment 1 and 2 using deverbal nouns and verbs respectively.

Experiment 1

While sharing the same stock of root morphemes with verbs, deverbal nouns draw on a specific set of word patterns which distinguishes them not only from verbs but also from the closed class of primitive nouns (Bohas and Guillaume, 1984). The purpose of this experiment is to investigate the time course of word pattern and root effects as opposed to semantic and form effects during the processing of deverbal nouns. To do this we used masked priming with four prime-display durations to assess priming between pairs of deverbal nouns which share either a word pattern, a root, or a non-structural orthographic or semantic relationship.

Method

Participants

A group of 138 volunteers aged 16 to 20 took part in the experiment. They were pupils at the high school of Tataouine in South Tunisia, and used MSA on regular basis.

Material

The prime and target pairs used fell into one of 6 conditions each of which comprised 24 pairs. In Condition 1, which we will refer to as [+WP], the prime and target share a word pattern (e.g., حارس-خالد [xaalid]-[aaris] "eternal"-guardian"). To control for the vocalic and consonantal overlap underlying the prime and target pairs in condition 1, Condition 2, [+Orth1] is an orthographic control condition matching the form overlap (primarily in shared vowels) of the word pattern pairs (e.g., سحابة-طلاق [sa aaba]-[t alaaq] "cloud"-"divorce"). In Condition 3, labeled [+R +S], the prime and target pairs share a root morpheme and a transparent semantic relationship (e.g., رئيس-رئاسة [ri aasa]-[ra iis] "presidency"-"president"). This is in contrast to Condition 4, labeled [+R –S], where the prime and target share a root but their semantic relationship is opaque (e.g., شرطة-شرط [art]-[urt a] "condition"-"police"). Con-dition 5, labeled [+Orth 2], is the orthographic control for the two conditions sharing a root (e.g., ثواب-ثابت [aabit]-[awaab] "firm"-"award"). The difference from [+Orth1] is that whereas the orthographic overlap in the latter relates to the shared vowels between prime and target, here overlap is specified solely in terms of shared consonants. Since vowels are not normally written in MSA (unless long), the form overlap in the [+Orth1] pairs is orthographically implicit, but fully explicit in the [+Orth2] pairs. Condition 6, [-R +S], consists of semantically but not morphologically related pairs (e.g., حرب-قتال [qitaal]-[arb] "fight"-"war"). Each of the related prime words was matched to an unrelated control prime. A similar number of pseudo-word-word pairs was constructed in such a way as to echo the form overlap between the word-word pairs.

Design and Procedure

Two versions were constructed such that all the targets appeared only once in each version, half

preceded by a related prime and half by an unrelated prime. Each trial consisted of three visual events. The first was a forward pattern mask, in the form of a sequence of 28 vertical lines in a 30-point traditional Arabic font size. The second event was a prime word written without diacritics in the same font but at 24 points. Four SOA's corresponding to a prime display duration of 32, 48, 64 and 80 ms were used. The third event was a target word or non-word written without diacritics in a 34 point font size. The larger font size of the target was used because MSA does not have the lower-case upper-case distinction. The stimulus words and non-words were written in the usual unvowelled script. Thirty two participants were assigned to the first SOA, forty to the second, thirty six to the third, and thirty to the fourth SOA. Participants were asked to make lexical decision about the target by pressing a "YES" or a "NO" key.

Results and discussion

Figure 1 plots the amount of priming (target RT when preceded by an unrelated prime minus target RT when preceded by a related prime) across condition and SOA.

Targets preceded by a prime with which they have a common word pattern were significantly facilitated only at SOA's 48 and 64 ms. The orthographic control for the word-pattern pairs [+Orth1] showed signs of priming at SOA 32, but no effects at any later SOA's, indicating that the word-pattern effects at SOA 48 and 64 are unlikely to be form-based. Word pattern priming in MSA deverbal nouns contrasts with the lack of word pattern priming in Hebrew nouns (Frost et al., 1997). This difference may be traced back to the differences underlying the word pairs making up the [+WP] condition in the present study and the Hebrew word pairs used in the same condition by Frost at al., (1997). In this study, as in our original study where we first report word pattern priming in Arabic nominal forms, we made a distinction between the syntactic meaning of the word pattern and its phonological structure (Boudelaa & Marslen-Wilson, 2000). Two surface forms may have a word pattern with the same surface phonological structure but with quite different syntactic meanings. For example, the pair قرود-نزول [quruud]-[nuzuul] ("monkeys"-"going down"), share the phonological structure of the word pattern, which is CVCVVC in both, but diverge with respect to its syntactic meaning. The word pattern has a "plural" meaning in [quruud], but a "singular deverbal noun" meaning in [nuzuul]. When we compared priming between deverbal nouns sharing both the syntactic meaning and the phonological structure of the word pattern with priming between nouns sharing only the phonological structure of the word pattern, it was only the former type of word pairs that yielded significant priming. The word pattern priming in the present experiment replicates our initial finding of reliable facilitation between deverbal nouns sharing the phonological as well as the syntactic meaning of the word pattern.

Turning to the two root conditions, [+R+S] and [+R-S], there was robust priming, at a constant level, across all four SOA's. In the [+Orth2] condition, where the prime and target shared a form overlap that mimicked the consonantal overlap in the root pairs, a facilitatory effect emerges at SOA 80 ms.

Similarly, significant facilitation emerges only at SOA 80 ms in the [-R+S] condition, where there is only a semantic relationship between prime and target. It is interesting to note that while the effects of morphology (here word pattern and root effects) are clearly distinct from orthographic and semantic effects at SOA's 32, 48 and 64 ms, the distinction between morphology-based and form-based and meaning-based effects dissipates at SOA 80. As can be seen in Figure 1 above, at SOA 80 the facilitation observed in the [+R +S] and the [+R-S] conditions is no longer different from that found in the [+Orth2] or the [-R+S] conditions.

A further point relates to the differences in priming between the two orthographic conditions. In [+Orth1], where the form overlap between prime and target is vowel-based, facilitation occurs only at the shortest SOA. In [+Orth2] where form overlap is consonantal in nature, significant priming takes place only at the longest SOA. These differential effects can be

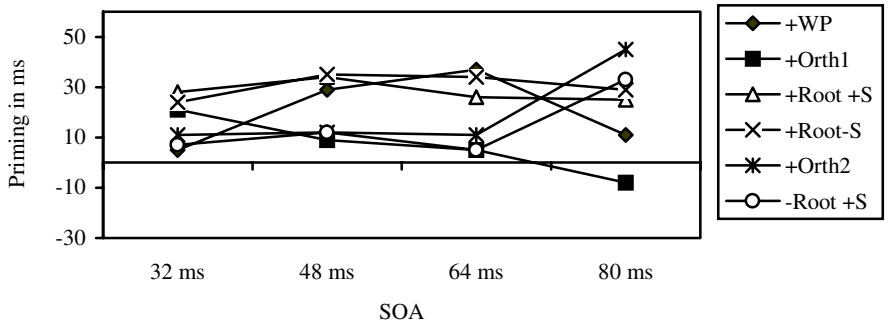


Figure (1): Priming in deverbal nouns as a function of relatedness and SOA

understood in the light of the different functional factors underlying word pattern and root priming – especially, as here, in the written modality, where consonantal information dominates the overt orthographic content of prime and target. The fact that root priming has a more precocious time course than word pattern priming suggests that the language processor monitors for the root consonants in order to access the meaning of the form at hand. Hence word pairs having a non-structural consonantal overlap may act as competitors early on in processing. Conversely, the slightly later onset of word pattern priming suggests that this unit comes into play once a root unit has been converged on. Since a root unit should be successfully extracted very quickly if meaning is to be accessed at all, word pairs sharing vocalic overlap need not compete with each other at the early stages of processing, hence the priming observed in the [+Orth1] condition.

In sum, the results of this experiment suggest that morphological effects take precedence over form driven and meaning driven effects. More importantly perhaps, from the perspective of Semitic morphology, both word pattern and root morphemes are actively used very early in processing. However, the effects of word patterns are rather transient yielding facilitation only over two SOA's, while the effects of the root are more durable (Deutsch et al., 2000). Arguably the qualitative difference underlying the effects of word patterns and roots reflects the difference between the information conveyed by these two units. Word patterns convey information that is syntactic and phonological in nature (Holes, 1995), whereas roots convey semantic information. The more consistent effects of roots by contrast to word patterns suggest that the information conveyed by the root is critically used throughout processing while that conveyed by the word pattern is only transiently salient.

Experiment 2

As we noted earlier, a primary division in Arabic morphology is between the sets of deverbal nouns and the verbs themselves. While roots are the same across deverbal nouns and verbs, word patterns are the unit that distinguishes these two syntactic categories. Accordingly, Experiment 2 was designed to ask three questions: First, will word patterns have similar transient priming effects in verbs as they do in nouns? Second, will root morphemes yield the same consistent priming effects in verbs? Third, are the effects of these two morphological units distinct from the effects produced by orthography and semantics?

Method

Participants

Another group of 108 participants from the same age group and linguistic background as those in Experiment 1 took part in this experiment.

Material and design

The design was analogous to that used in Experiment1. The material consisted of prime and target verb forms which made up 6 experimental conditions with 24 pairs each. In Condition 1, [+WP], the prime and target share a word pattern (e.g., أبلغ-أحرز [a raza]-[abla a] "obtain"-"inform"). Condition 2, [+Orth 1], was an orthographic control for the form overlap in condition 1 (e.g., أنكر-لجنة [la natun]-[ankara] "committee"-"deny"). In condition 3, [+R +S], prime and target share a root morpheme and a transparent semantic relationship (e.g., أحرق-احترق [i taraqa]-[a raqa] "get burned"-"burn"), while in condition 4, [+R –S], they share a root but have an opaque semantic relationship (e.g., تقدم-تقادم [taqaadama]-[taqaddama] "get old"-"progress"). As in Experiment 1, Condition 5, [+Orth 2], is the orthographic control for conditions 3 and 4 (e.g., بلل-بلع [bala a]-[ballala] "swallow"-"soak"). Condition 6, [-R +S], tests for purely semantic effects (e.g., تأكد -أيقن [ayqana]-[ta akkada] "ascertain"-"make sure"). Each of the related primes, across the 6 conditions, was matched as closely as possible to a control prime that shared no relationship with the target. A similar number of pseudo-word-word pairs was constructed in such a way as to echo the form overlap between the word-word pairs.

Procedure

The procedure was the same as in Experiment 1.

Results and discussion

Figure 2 shows net priming effects for the six experimental conditions across the four SOA's.

The results are even more clear cut than for the deverbal nouns. The effects of word pattern morphemes is again strong but highly transient, yielding significant priming at SOA 48 only. Its matched orthographic

control, [+Orth1] condition shows a marginal 12 ms facilitation at SOA 32 ms but no effects thereafter, suggesting that word pattern effects are genuine morphological effects that are not amenable to a form-based account. As regards roots, significant priming effects are observed across all four SOA's in both the [+R+S] and the [+R-S] conditions. Orthographic and semantic effects, as illustrated by the [+Orth2] and the [-R+S] conditions, again emerge only at SOA 80. This confirms that root morpheme effects and form-driven and meaning-driven effects have different time courses, the latter two effects taking more time to build up.

These results suggest that in MSA, morphological processing has a different locus from form-based and meaning-based processing. Moreover, verb word patterns, like deverbal noun word patterns, play a highly significant, though transient role during processing. Roots, by contrast, give rise to an evenly distributed pattern of facilitation across SOA's. The different time courses of word pattern and root processing observed in this experiment and in the previous one, suggest that the language processor uses the information conveyed by these two units in different ways and at different points in the internal process of linguistically interpreting a written form.

Conclusion

We have reported two experiments aimed at assessing the time course of morphological, orthographic and semantic effects. In so far as Semitic languages are concerned, there are at least three ways in which morphological effects can be said to be clearly distinct from orthographic and semantic effects:
First, word pattern morphemes, which are non-semantic in nature, yield significant priming, while their matched orthographic controls either do not yield any priming at all, as in Experiment 2, or do so much less reliably, as in Experiment 1. Second, root morphemes play a role irrespective of semantic transparency, with surface forms giving rise to reliable and significant priming as long as they share a morphemic unit. Third, morphological effects occur prior to orthographic and semantic effects and have a longer time-course - at least as far as the root is concerned. In the context of Semitic morphology, this is the first demonstration that word pattern and root morphemes have overlapping but different processing time courses. This state of affairs is a direct consequence of the kind of information that word patterns and roots convey. The reliable and long-lived root priming effects reflect the fact that lexical interpretation and integration of Arabic surface forms relies primarily on this unit. The transient word pattern priming effects point to the fact that this unit is the focus of the lexical mapping process only in so far as a consonantal root unit can be successfully extracted. Evidence supporting this comes from the finding that no word pattern priming is found with pseudowords consisting of existing word patterns and a non-existing root, while root priming is found in pseudo words consisting of an illegal combination of an existing word pattern and an existing root (Frost et al., 1997). Functionally, the results point to the conclusion that morphemic units that are non-linear and abstract are able to govern lexical access and lexical representation.

Turning to the orthographic effects observed in this study, it seems that in MSA, and arguably in other Semitic languages as well, vowels and consonants have a different status. This is clear from the differential priming yielded by word patterns and consonants on the one hand, and by the different loci of orthographic priming in the [+Orth1] and [+Orth2] on the other. Remember that when
orthographic overlap is defined in terms of shared vowels across primes and targets as in the [+Orth1] condition, facilitation is early and transient. By contrast, when it is defined in terms of the consonants shared by prime and target as in the [+Orth2] condition, facilitation is late and robust. Taken together, these results suggest that proximity in the Arabic lexical space is sensitive to similarity in vowels and to similarity in consonants, and that early on in processing lexical competition is initiated on the basis of the consonantal component of the surface form but not its vocalic component.

It is worth noting that in English, where the vowel-consonant distinction is not morphemic, and under experimental conditions that are most similar to ours, that is at 32 ms SOA (Feldman, 2000) and at 43 ms (Rastle et al., 2000), no orthographic effects are obtained. Nonetheless, overall orthographic effects in English tend to be facilitatory early in processing and inhibitory later on. In other words, English orthographic priming is more in keeping with the priming profile we observe in our [+Orth1] condition, and the mirror image of what we observe in our [+Orth2] condition. The emergence of relatively late form effects as evidenced by the priming in [+Orth2] at SOA 80 can perhaps be accommodated within a model where knowledge about the semantic and formal attributes of the input, be they phonological or orthographic, are at the same level in the perceptual system and are computed in parallel, rather than having the form computed prior to accessing the lexicon (Gaskell & Marslen-Wilson, 1998).

From the perspective of a general theory of morphological processing and representation, the current results put new constraints on how to account for morphological effects. For example, while connectionist models, as they now stand, may be able to account for morphological priming in the absence of synchronic semantic links between prime and target (Plaut & Gonnerman, 2000), it is less clear how word patterns, which are non-semantic units in essence, would be predicted to generate priming within such a framework. Nevertheless, it remains important to persevere with models which offer explicit and quantitative predictions about behavior.

Acknowledgement

We thank Mike Ford for his assistance in the preparation of the two experiments and Abdallah Megbli, headmaster of the High School in Tataouine, Tunisia for his generous help in providing testing facilities and access to participants for testing. We also thank Matt Davies, for helpful comments on an earlier version of this work.

References

Bohas, G. & Guillaume, J.-P. (1984) *Etudes des Th ories des grammairiens Arabes: I Morphologie et Phonologie*: Damas, Syria.

Boudelaa, S., & W. D. Marslen-Wilson (2000). Non-concatenative morphemes in language processing: Evidence from Modern Standard Arabic. *Proceedings of the Workshop on Spoken Word Access Processes, 1*, 23-26, Nijmegen, Netherlands.

Deutsch, A., Frost, R., & Forster, K. (1998). Verbs and nouns are organized and accessed differently in the mental lexicon: Evidence from Hebrew. *Journal of Experimental Psychology: Learning Memory, and Cognition, 24*, 1238-1255.

Deutsch, A., Frost, R., Pollatsek, A, & Rayner, K. (2000). Early morphological effects in word recognition in Hebrew: Evidence from parafoveal preview benefit: *Language and Cognitive Processes, 15*, 487-506.

Feldman, L. B. (2000). Are morphological effects distinguishable from the effects of shared meaning and shared form. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 6*, 1431-1444.

Forster, K. I. & Azuma, T0 (2000). Masked priming for prefixed words with bound stems: Does submit prime permit? *Language and Cognitive Processes,* 15, 539-561.

Frost, R., Forster, K. I., & Deutsch, A. (1997). What can we learn from the morphology of Hebrew? A masked priming investigation of morphological representation. *Journal of Experimental Psychology: Learning, Memory and Cognition, 23*, 829-856.

Gaskell, M. G, & Marslen-Wilson, W. D. (1998). Mechanisms of phonological inference in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 380-396.

Holes, C. (1995). *Modern Arabic*. Longman, London-NY.

Perea, M., Gotor, A., Rosa, E., & Algarabel, S. (1995, Nov). Time course of semantic activation for different prime prime-target relationships in the lexical decision task.
36th Annual meeting of the Psychonomic Society, Los Angles, California.

Plaut, D. & Gonnerman, L. M. (2000). Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes,* 15, 445-485.

Rastle, K., Davis, M. H., Marslen-Wilson, W. D., & Tyler, L. K. (2000) Morphological and semantic effects in visual word recognition: A time-course study. *Language and Cognitive Processes, 15*, 507-537.

Sereno, J. A. (1991). Graphemic, associative, and syntactic priming effects at brief stimulus onset asynchrony in lexical decision and naming. *Journal of Experimental Psychology: Learning, Memory and Cognition, 17*, 459-477.

Versteegh, K. (1997). *The Arabic language*. Edinburgh University Press

Wright, W. (1995). *A Grammar of the Arabic Language*, Cambridge University Press.

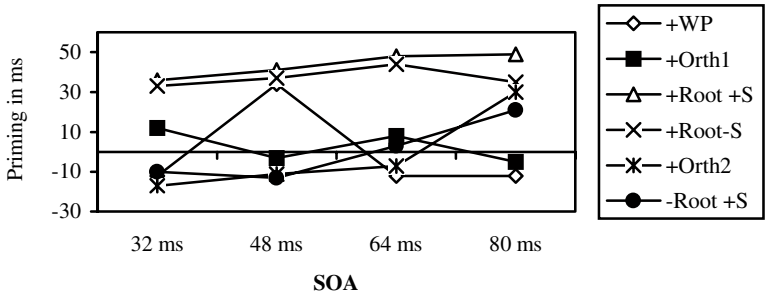


Figure (2): Priming in verbs as a function of relatedness and SOA

Reference-point Reasoning and Comparison Asymmetries

Brian F. Bowdle (bbowdle@indiana.edu)
Department of Psychology, Indiana University
1101 East Tenth Street
Bloomington, IN 47405 USA

Douglas L. Medin (medin@nwu.edu)
Department of Psychology, Northwestern University
2029 Sheridan Road
Evanston, IL 60208 USA

Abstract

Comparison asymmetries are most often explained in terms of underlying asymmetries in the perceived similarity of the comparison items, which in turn are seen as arising from the differential weighting of distinctive features of the target and base representations. In two experiments, we fail to confirm the predictions of the standard account. Rather, comparison asymmetries seem to follow from two general principles. First, certain items act as cognitive reference points that other, less prominent category members are located in terms of or assimilated to. And second, the target and base terms of a comparison play different semantic roles, with the target acting as the figure and the base acting as the ground.

Introduction

The notion that similarity is a symmetric relation is highly intuitive. After all, if one claims that limes are similar to lemons, this would seem to entail that lemons are also similar to limes. This notion is further supported by the observation that many comparisons can be stated either directionally, as in Limes are similar to lemons, or non-directionally and reciprocally, as in Lemons and limes are similar to each other. Nevertheless, comparisons often behave asymmetrically. For example, Tversky (1977) showed that people frequently prefer one direction of comparison (e.g., North Korea is similar to China) over the other (e.g., China is similar to North Korea). Such asymmetries are even more pronounced in metaphors and similes, for which only one direction of comparison may be meaningful. For example, whereas Time is like a river is an informative statement, A river is like time is nonsensical. The general observation is that, whenever two items differ in prominence due to such factors as familiarity, salience, or concreteness, the less prominent item is compared to the more prominent item.

What is the source of these comparison asymmetries? That is, given that two items are recognized as being similar, why should one direction of comparison be more natural and meaningful than the other? Clearly, the answer to this question is important to any psychologically plausible model of comparison. Indeed, the existence of comparison asymmetries has been used to argue for and against different theories of similarity (e.g., Tversky, 1977) and metaphor comprehension (e.g., Glucksberg & Keysar, 1990; Ortony, 1979). In this paper we evaluate two different accounts of the cognitive factors underlying comparison asymmetries.

The Standard Account

Comparison asymmetries are most commonly explained in terms of Tversky's (1977) contrast model of similarity which predicts that, under certain circumstances, the similarity of item a to item b will actually seem greater than that of item b to item a. According to the contrast model, the perceived similarity of item a to item b, $s(a,b)$, is given by

$$s(a,b) = qf(A \cap B) - af(A - B) - bf(B - A)$$

where A and B are the features of a and b, f is a measure of salience, and q, a, and b are weights assigned to the feature sets. The basic idea is that the similarity of two items increases as a function of their common features and decreases as a function of their distinctive features. Asymmetries in the similarity of two items are predicted in terms of the focusing hypothesis: Because the target (first term) of a directional comparison is the subject of the statement, it will receive more attention than the base (second term). This means that the distinctive features of the target are weighted more heavily than those of the base – that is, $a > b$. Thus, the similarity of a to b will seem greater than that of b to a whenever b possesses the larger or more salient set of distinctive features.

Consistent with the contrast model, asymmetric similarity ratings have been obtained in a wide range of stimulus domains, such that less prominent items are seen as being more similar to more prominent items (e.g., Bartlett & Dowling, 1988; Holyoak & Gordon, 1983; Ortony, Vondruska, Foss, & Jones, 1985; Tversky, 1977). But how does this explain the fact that people typically prefer one direction of comparison between two items over the other? The standard answer is that, when interpreting a similarity comparison, the hearer seeks to maximize the similarity of the items. In other words, people prefer North Korea is similar to

China over the reverse direction of comparison precisely because North Korea is judged as being more similar to China than the reverse – presumably reflecting differences in the featural complexity of the two items. In support of this position, both Tversky (1977) and Ortony et al. (1985) found that items in the preferred comparison order typically received higher similarity ratings than the same items in the non-preferred order.

To summarize, the standard account of comparison asymmetries makes two claims. First, comparison asymmetries reflect underlying asymmetries in the perceived similarity of the items. And second, these underlying asymmetries are due to attentional factors, such that the distinctive features of the target are weighted more heavily than the distinctive features or spatial density of the base.

Cognitive Reference Points

Although the contrast model has been widely adopted, there is an alternative explanation of comparison asymmetries – namely, that such asymmetries follow from principles of reference point reasoning (e.g., Gleitman, Gleitman, Miller, & Ostrin, 1997; Roese, Sherman, & Hur, 1998; Rosch, 1975; Shen, 1989). One of the central claims of this position is that certain highly prominent items act as cognitive reference points that other items are seen in relation to. Some well-known examples of cognitive reference points are prototypes and ideals, which may be used to understand less prominent category members (Rosch, 1975), and the self concept, which serves as a habitual landmark in social judgments (e.g., Holyoak & Gordon, 1983; Srull & Gaelick, 1983). The basic idea is that many domains of knowledge are at least partially structured in terms of a small number of reference items.

Of course, the claim that non-reference (or deviant) items are seen in relation to reference items raises the question of what is meant by "seen in relation to." One way in which this relationship may manifest itself is conceptual location: Cognitive reference points provide landmarks that can be used to better specify the location of deviant items in a semantic or perceptual space. By doing so, reference items lend stability to the representations of deviant items. For example, it may be easier to conceptualize and reason with non-standard quantities (e.g., a length of two feet and nine inches) in terms of certain standards of measurement (e.g., a length of one yard). The beneficial use of reference items as landmarks for locating deviant items has been demonstrated in several studies of magnitude comparisons, where pairs of deviant items were discriminated with greater speed and accuracy when they were in the vicinity of a cognitive reference point (e.g., Holyoak & Mah, 1982; Hutchinson & Lockhead, 1977; te Linde & Paivio, 1979).

In addition to conceptual location, there is a second and more complex way in which deviant items may be seen in relation to cognitive reference points – namely, conceptual assimilation. The idea here is that deviant items are more easily assimilated to reference items than the reverse (e.g., Bowdle & Gentner, 1997; Rosch, 1975; Shen, 1989). Such assimilation effects have been obtained in numerous studies. For example, people are more likely to project new properties from prototypical category members to less prominent members than vice versa (Rips, 1975), and are more willing to make inferences and predictions about others based on the self than vice versa (e.g., Kunda & Nisbett, 1988; McFarland & Miller, 1990). Whenever such assimilation occurs, the representation of the deviant item is changed to make it more concordant with that of the reference item.

The above discussion of the functions of cognitive reference points suggests that, even prior to being placed in a comparison, there is a directional or asymmetric relationship between two items whenever one makes a better cognitive reference point than the other. But how does this translate into preferred comparison orders? An answer commonly given by reference point models is that the target and base terms of a comparison play different semantic roles, which specify the placement of deviant and reference items in the comparison frame.

It has been claimed that items in the subject and complement positions of many sentence types are assigned the roles of figure and ground, respectively (Gleitman et al., 1997; Langacker, 1990; Talmy, 1978). The figure is characterized as a moving or conceptually movable object whose site or path is the issue of interest. In contrast, the ground is characterized as a stationary landmark with respect to which the figure's site or path is defined. Thus, whichever item makes a more natural cognitive reference point will be the preferred ground of the sentence. In directional comparisons, this predicts that deviant items should be placed in the target position and reference items in the base position.

Perhaps the most notable distinction between the standard account of comparison asymmetries and the reference point account is that the latter does not rely on the notion of underlying asymmetries in the perceived similarity of the comparison items. That is, one does not have to judge whether item a seems more similar to item b or item b seems more similar to item a in order to determine their preferred ordering. Rather, comparison asymmetries reflect the fact that deviant items are more concordant with the semantic constrains of the target position, and reference items with the semantic constraints of the base position. Simply put, using a cognitive reference point as the base of a directional comparison results in a more natural and informative statement.

Comparing the Positions

Both the standard account and the reference point account are able to explain many of the comparison

asymmetries that have been observed in the literature, albeit using different mechanisms. In the present study, we sought to address an important limitation of existing research in this area. Specifically, the available evidence almost exclusively involves asymmetries in similarity comparisons, for which the two accounts make essentially the same predictions concerning which direction of comparison should be preferred. If one turns to consider the relationship between similarity and difference comparisons, however, then the two accounts can be shown to make distinct predictions.

According to the standard account, people prefer the direction of a similarity comparison that maximizes the perceived similarity of the target to the base. By analogy, then, people should also prefer the direction of a difference comparison that maximizes the perceived difference of the target from the base. This suggests that comparison asymmetries should go in opposite directions for similarity and difference statements, as asymmetries in similarity and difference ratings tend to be inversely related (Tversky, 1977). For example, if North Korea seems more similar to China than the reverse, then China will seem more different from North Korea than the reverse. Therefore, people should not only prefer North Korea is similar to China over China is similar to North Korea, they should also prefer China is different from North Korea over North Korea is different from China. In both cases, the preferred direction of comparison maximizes the value of the dimension specified by the comparison predicate.

In contrast to the standard account, the reference point account states that people simply prefer the direction of comparison that uses the better cognitive reference point as the ground, because this ordering maximizes the informativity of the statement. Given that the position of figure and ground in a statement should not be affected by the particular comparison predicate, the preferred direction of comparison between two items should place reference items in the base position for both similarity and difference statements. Thus, if people prefer North Korea is similar to China over the reverse, then they should also prefer North Korea is different from China over the reverse.

In addition to making different predictions about the direction of comparison asymmetries for similarity and difference statements, the standard and reference point accounts also make different predictions about the relative magnitude of such asymmetries. According to Tversky (1977), difference comparisons will tend to place more weight on the distinctive feature sets than will similarity comparisons. Because the standard account derives asymmetries from distinctive features, this means that difference comparisons should be more asymmetric than similarity comparisons. In contrast, the reference point account suggests precisely the opposite – similarity comparisons should be more asymmetric than difference comparisons. Although the use of reference items to specify the location of deviant items is presumably equally important in similarity and difference statements, conceptual assimilation of deviant items to reference items should be more likely to occur in similarity statements. As noted by a number of theorists, informative similarity comparisons do not merely point out obvious commonalities; rather, they highlight nonobvious commonalities, and promote the creation of new ones through processes such as inference projection (e.g., Bowdle & Gentner, 1997; Medin et al., 1993). While less work has been done concerning the communicative functions of difference comparisons, it is reasonable to assume that difference comparisons are less likely to invite such modes of conceptual assimilation. This is because difference comparisons serve more to suggest differences between items than to suggest commonalities. Thus, although there should be a general preference for comparing deviant items to reference items, the utility of doing so should be greater for similarity statements than for difference statements.

Experiment 1

In Experiment 1, we tested the central predictions of the standard and reference point accounts concerning comparison asymmetries. Subjects were given directional similarity or difference comparisons, each of which contained a less prominent (deviant) item and a more prominent (reference) item. All comparisons were presented in both possible orders – with the reference item in the base position (e.g., A zebra is similar to/different from a horse) or in the target position (e.g., A horse is similar to/different from a zebra). For convenience, we will refer to statements with the first ordering of items as forward comparisons, and statements with the second ordering of items as reverse comparisons. For each comparison, subjects were asked to indicate the strength of their preference for one direction of comparison over the other. Again, the standard account predicts that comparison asymmetries should go in opposite directions for similarity and difference statements, and should be stronger for difference statements. In contrast, the reference point account predicts that comparison asymmetries should go in the same direction for similarity and difference statements, and should be stronger for similarity statements.

Method

Subjects. Forty Northwestern University undergraduates participated in partial fulfillment of a course requirement.

Materials and Design. Each subject received 32 directional comparisons between a less prominent (deviant) item and a more prominent (reference) item. (The relative prominence of each item was initially determined by the authors and then confirmed by two judges.) To ensure generality, the 32 comparisons involved eight categories of items: animals (e.g., zebra – horse), artifacts (e.g., motel – hotel), colors (e.g., tan – brown), countries (e.g., North Korea – China), emotions

(e.g., admiration – love), famous individuals (e.g., Saddam Hussein – Adolf Hitler), measurements (e.g., $105.00 – $100.00), and occupations (e.g., dentist – surgeon).

Half of the subjects received all 32 comparisons as similarity statements (e.g., A zebra is similar to a horse), and half as difference statements (e.g., A zebra is different from a horse). Subjects saw each statement in both forward and reverse directions, with the two directions separated by a six-point numerical scale. The order of presentation of the two directions (forward first versus reverse first) was counterbalanced within and between subjects.

Procedure. Each subject was given a booklet containing the 32 pairs of comparison statements in a random order. Subjects indicated which direction of comparison they felt was "stronger, more sensible, or more natural" for each pair by circling a number on the six-point scale. They were told that the more strongly they preferred the direction on the left, the closer their answer should be to 1, and the more strongly they preferred the direction on the right, the closer their answer should be to 6.

Results and Discussion

All directional preference ratings were transformed so that higher numbers indicated a preference for forward comparisons over reverse comparisons. For similarity statements, the directional preference ($M = 4.77$, $SD = 0.39$) was significantly above the scale midpoint (3.5) by both subjects and items, $t_s(19) = 14.66$, $p < .001$ and $t_i(31) = 19.44$, $p < .001$. For difference statements, the directional preference ($M = 4.03$, $SD = 0.57$) was also significantly above the scale midpoint, $t_s(19) = 4.14$, $p < .001$ and $t_i(31) = 6.41$, $p < .001$. Thus, subjects consistently preferred comparing deviant items to reference items in both similarity and difference statements. This is consistent with the reference point account of comparison asymmetries: People prefer the direction of comparison that places the better cognitive reference point in the base position, regardless of the particular comparison predicate used.

Turning to the relative magnitudes of the comparison asymmetries, the preference for the forward direction of comparison was higher for similarity statements than for difference statements, $t_s(38) = 4.83$, $p < .001$ and $t_i(31) = 10.31$, $p < .001$. Again, this is as predicted by the reference point account: Because similarity statements are likely to elicit a greater degree of conceptual assimilation than difference statements, reference point effects should be stronger in similarity statements.

Asymmetries in Similarity and Difference Ratings

Contrary to the claims of the standard account, the results of Experiment 1 suggest that comparison asymmetries are not due to underlying asymmetries in the perceived similarity or difference of the comparison items. If this were the case, then – assuming that hearers seek to maximize the value of the dimension specified by the comparison predicate – comparison asymmetries should have gone in opposite directions for similarity and difference statements. But how, then, does one explain the fact that comparison asymmetries are typically associated with asymmetries in similarity and difference ratings (e.g., Ortony et al., 1985; Tversky, 1977)? We suggest that such ratings asymmetries might also be due to reference point reasoning.

According to the reference point account, the target and base terms of a directional comparison play different semantic roles, with the target acting as the figure and the base acting as the ground. Thus, information flows directionally from the base to the target, as when the base is used to generate new inferences about the target. Assuming that deviant items are more easily assimilated to reference items than the reverse, this means that assigning the reference item to the base position (forward comparisons) should result in a greater degree of conceptual assimilation than assigning it to the target position (reverse comparisons). Therefore, forward comparisons should elicit higher similarity ratings – and lower difference ratings – than reverse comparisons.

This explanation of ratings asymmetries is radically different from that offered by Tversky's (1977) contrast model. In this model, the representations of the comparison items are assumed to remain static, and asymmetries are simply due to attentional factors. On the reference point view, however, the representations of deviant items may shift towards those of reference items, thereby making the items more similar. This view is, in fact, consistent with a fair amount of evidence. Indeed, asymmetries in conceptual assimilation are often associated with asymmetries in similarity ratings. For example, people not only make more inferences and predictions about others based on the self than vice versa (e.g., Kunda & Nisbett, 1988; McFarland & Miller, 1990), they also rate others as being more similar to the self than vice versa (e.g., Catrambone, Beike, & Niedenthal, 1996; Holyoak & Gordon, 1983; Srull & Gaelick, 1983). We propose that the latter effect may be largely due to the former – projecting novel information from the self to others will make others seem more similar to the self.

In sum, the reference point account can explain asymmetries in similarity and difference judgments, and in fact predicts the same directionalities as the standard account. As was the case for comparison asymmetries, however, these approaches make different predictions about the relative magnitude of asymmetries in similarity and difference ratings. According to the standard account, difference comparisons will tend to place more weight on the comparison items' distinctive feature sets than will similarity comparisons. Because the standard account derives asymmetries from precisely these stimulus properties, this predicts that directional difference ratings should be more asymmetric than directional similarity ratings. According to the reference point account, however, this pattern of results

should not hold. This is because conceptual assimilation is more likely to occur in similarity comparisons. Assuming that conceptual assimilation is in fact a primary source of ratings asymmetries, then, directional similarity ratings should be more asymmetric than directional difference ratings.

Experiment 2

In Experiment 2, subjects were given the same directional comparisons used in Experiment 1, and rated either the similarity or the difference of both the deviant item to the reference item (e.g., How similar is a zebra to a horse?) and the reference item to the deviant item (e.g., How similar is a horse to a zebra?). Again, the standard account predicts that difference judgments should be more asymmetric, whereas the reference point account predicts that similarity judgments should be more asymmetric. We also gave a second group of subjects nondirectional versions these comparison questions (e.g., How similar are a zebra and a horse? or How similar are a horse and a zebra?). That is, these subjects were asked to rate either the similarity of or the difference between the two items without any specification of which item was the target and which was the base.

The inclusion of the nondirectional ratings condition was inspired by Catrambone et al. (1996), who argued that if the more prominent of two comparison items serves as a cognitive reference point for understanding the other item, then it should act as the implicit base of a nondirectional comparison. That is, nondirectional comparisons should be mentally translated into forward comparisons, in which the deviant item is directionally compared to the reference item. Supporting this claim, Catrambone et al. found that nondirectional similarity comparisons were rated as expressing the same degree of similarity as forward similarity comparisons, and a higher degree of similarity than reverse similarity comparisons. In the present experiment, we sought to replicate this finding for similarity comparisons, and extend it to difference comparisons. If both nondirectional similarity and difference ratings are closer to forward than reverse ratings, then this would further support the claim that asymmetries are due to reference point reasoning.

Method

Subjects. Eighty Northwestern University undergraduates served as paid subjects.

Materials and Design. Half of the subjects were assigned to the directional ratings condition, and half to the nondirectional ratings condition. In the directional condition, subjects received all 32 directional comparisons used in Experiment 1. Half of the subjects in this condition were asked to rate the similarity of the comparison items, and half the difference between the comparison items. For each comparison, subjects gave ratings for both the forward direction and the reverse direction. The order of presentation of the two directions was counterbalanced within and between subjects.

In the nondirectional ratings condition, subjects received nondirectional versions of the 32 comparison statements. As in the directional condition, half of the subjects were asked to rate the similarity of the comparison items, and half the difference between the comparison items. Because nondirectional comparisons lack target and base terms, however, subjects gave only one rating per comparison in this condition. The order of presentation of the deviant and reference items in a comparison (e.g., How similar are a zebra and a horse? versus How similar are a horse and a zebra?) was counterbalanced within and between subjects.

Procedure. Each subject was given a booklet containing the 32 comparison statements in a random order. Subjects gave similarity or difference ratings by circling a number on a 20-point scale below each comparison. For similarity ratings, the low end of the scale was labeled "not at all similar" and the high end "very similar". For difference ratings, the low end was labeled "not at all different" and the high end "very different".

Results and Discussion

Focusing first on the directional ratings, subjects gave higher similarity ratings to forward comparisons ($M = 11.02$, $SD = 2.44$) than to reverse comparisons ($M = 9.84$, $SD = 2.69$), $t_s(19) = 3.93$, $p < .001$ and $t_i(31) = 6.57$, $p < .001$. Likewise, subjects gave higher difference ratings to reverse comparisons ($M = 13.12$, $SD = 2.26$) than to forward comparisons ($M = 12.44$, $SD = 2.35$), $t_s(19) = 3.29$, $p < .005$ and $t_i(31) = 3.42$, $p < .005$. These results are consistent with both the standard account and the reference point account. More critically, however, the directional similarity ratings were more asymmetric than the directional difference ratings: The absolute mean difference in ratings between the forward and reverse comparisons was nearly twice as large for similarity comparisons ($M = 1.18$, $SD = 1.34$) as it was for difference comparisons ($M = 0.68$, $SD = 0.92$). This is only consistent with the reference point account, according to which conceptual assimilation will result in asymmetric similarity and difference ratings but is more likely to occur in similarity comparisons. However, this difference in the magnitude of the ratings asymmetries was only marginally significant by items, $t_i(31) = 1.91$, $p < .10$, and not by subjects, $t_s(38) = 1.38$, $p < .20$.

Turning now to consider the entire pattern of ratings, the nondirectional similarity ratings ($M = 11.42$, $SD = 2.71$) did not differ from forward similarity ratings, but were significantly larger than reverse similarity ratings, $t_s(38) = 2.23$, $p < .05$ and $t_i(31) = 6.41$, $p < .001$. This replicates the findings of Catrambone et al. (1996). Likewise, the nondirectional difference ratings ($M = 12.10$, $SD = 3.12$) did not differ from forward difference ratings, but were significantly smaller than reverse difference ratings by items, $t_i(31) = 4.25$, $p < .001$, but

not by subjects, $t_s(38) = 1.58, p < .20$. Thus, subjects seemed to interpret nondirectional similarity and difference comparisons as forward comparisons, in which the reference item played the implicit role of ground. This result cannot be explained by Tversky's (1977) contrast model, and further illustrates the centrality of reference point reasoning in comparisons.

Conclusions

Our findings suggest that asymmetries in similarity and difference comparisons cannot be explained in terms of the differential weighting of static representations. Rather, they seem to follow from two general principles. First, certain items act as cognitive reference points that other items are understood in terms of via conceptual location or conceptual assimilation. And second, the target and base terms of a comparison play different semantic roles – the base, acts as the ground, is used to understand the target, which acts as the figure. Thus, comparison asymmetries reflect the fact that deviant items are more concordant with the linguistic constrains of the target position, and reference items with the linguistic constraints of the base position. Directional comparisons are maximally informative when a cognitive reference point is used as the base. Further, this direction of comparison is most likely to result in higher similarity ratings – and lower difference ratings – due to the increased potential for conceptual assimilation. In sum, the comparison process would appear to be far more dynamic than is commonly assumed, with reference-point reasoning playing a prominent role in both similarity and difference.

Acknowledgments

We thank Dedre Gentner, Robert Goldstone, Steven Sherman, and Phillip Wolff for their comments and suggestions. We also thank Gina Davis, Elizabeth Frame, Jason Griffith, and Matthew Kinman for their help with data collection and coding.

References

Bartlett, J.C., & Dowling, W.J. (1988). Scale structure and similarity of melodies. Music Perception, 5, 285-315.

Bowdle, B.F., & Gentner, D. (1997). Informativity and asymmetry in comparisons. Cognitive Psychology, 34, 244-286.

Catrambone, R., Beike, D., & Niedenthal, P. (1996). Is the self-concept a habitual referent in judgments of similarity? Psychological Science, 7, 158-163.

Gleitman, L., Gleitman, H., Miller, C., & Ostrin, R. (1997). Similar, and similar concepts. Cognition, 58, 321-376.

Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. Psychological Review, 97, 3-18.

Holyoak, K.J., & Gordon, P.C. (1983). Social reference points. Journal of Personality and Social Psychology, 44, 881-887.

Holyoak, K.J., & Mah, W.A. (1982). Cognitive reference points in judgments of symbolic magnitude. Cognitive Psychology, 14, 328-352.

Hutchinson, J.W., & Lockhead, G.R. (1977). Similarity as distance: A structural principle for semantic memory. Journal of Experimental Psychology: Human Learning and Memory, 3, 660-678.

Kunda, Z., & Nisbett, R.E. (1988). Predicting individual evaluations from group evaluations and vice versa: Different patterns for self and other? Personality and Social Psychology Bulletin, 14, 326-334.

Langacker, R.W. (1990). Subjectification. Cognitive Linguistics, 1, 5-38.

McFarland, C., & Miller, D.T. (1990). Judgments of self-other similarity: Just like other people, only more so. Personality and Social Psychology Bulletin, 16, 475-484.

Medin, D.L., Goldstone, R.L., & Gentner, D. (1993). Respects for similarity. Psychological Review, 100, 254-278.

Ortony, A. (1979). Beyond literal similarity. Psychological Review, 86, 161-180.

Ortony, A., Vondruska, R.J., Foss, M.A., & Jones, L.E. (1985). Salience, similes, and the asymmetry of similarity. Journal of Memory and Language, 24, 569-594.

Rips, L.J. (1975). Inductive judgments about natural categories. Journal of Verbal Learning and Verbal Behavior, 14, 665-681.

Roese, N.J., Sherman, J.W., & Hur, T. (1998). Direction of comparison asymmetries in relational judgment: The role of linguistics norms. Social Cognition, 16, 353-362.

Rosch, E. (1975). Cognitive reference points. Cognitive Psychology, 7, 532-547.

Shen, Y. (1989). Symmetric and asymmetric comparisons. Poetics, 18, 517-536.

Srull, T.K., & Gaelick, L. (1983). General principles and individual differences in the self as a habitual reference point: An examination of self-other judgments of similarity. Social Cognition, 2, 108-121.

Talmy, L. (1978). Figure and ground in complex sentences. In J. Greenburg, C. Ferguson, & M. Moravcsik (Eds.), Universals of human language, vol. 4. Stanford: Stanford University.

Te Linde, J., & Paivio, A. (1979). A symbolic comparison of color similarity. Memory and Cognition, 7, 141-148.

Tversky, A. (1977). Features of similarity. Psychological Review, 84, 327-352.

Deference in Categorisation: Evidence for Essentialism?

Nick Braisby (N.R.Braisby@open.ac.uk)
Department of Psychology, The Open University, Walton Hall,
Milton Keynes, MK7 6AA, UK

Abstract

Many studies appear to show that categorization conforms to psychological essentialism (e.g., Gelman & Wellman, 1991). However, key implications of essentialism have not been scrutinized. These are that people's categorizations should shift as their knowledge of micro-structural properties shift, and that people should defer in their categorizations to appropriate experts. Three studies are reported. The first shows that even gross changes in genetic structure do not radically shift categorizations of living kinds. The second and third reveal a pattern of conditional deference to experts, coupled with systematic deference to non-experts. It is argued that these results point towards only a partial role for essentialism in explaining categorization, and a continuing role for theories that emphasize the importance of appearance and/or functional properties.

Introduction

Theories of concepts have shifted markedly over the past twenty years (Medin, 1989). Early theories were characterised by the view that similarity determines category membership, and that similarity-based models can account for a range of empirical evidence concerning human categorisation (e.g., Rosch, 1975; Rosch & Mervis, 1975). However, Murphy & Medin (1985), building on the earlier critiques of philosophers such as Goodman (1972), argued that similarity was a notion of weak explanatory value. Instead, they proposed that lay or common-sense theories were responsible for determining category membership.

There have since been many demonstrations of the influence of common-sense theories on category learning (e.g., Murphy, 1993; Kaplan & Murphy, 2000; Spalding & Murphy, 1996) and on categorisation itself (e.g., Keil, 1986, 1989; Rips, 1989). Nonetheless difficulties with theory-based accounts of categorisation remain. Margolis & Laurence (1999) and Fodor (1998) both point to the difficulties for theory-based views in accounting for error and ignorance.

A variant of the theory-based approach has also been proposed – psychological essentialism (Medin, 1989; Medin & Ortony, 1989). According to this, people tend to believe that objects have essences that are grounded in micro-structural properties (e.g., genetic properties for categorising organisms), and it is this belief that guides their categorisation, even though the belief may turn out to be false. This psychological essentialism differs from metaphysical essentialism (cf. Kripke, 1980; Putnam, 1975), which is the stronger doctrine that members of natural categories do in fact possess essences that determine their category membership.

Much evidence has been cited in support of psychological essentialism. Aside from earlier demonstrations from Carey (1985), Keil (1986, 1989) and Rips (1989) that similarity does not always determine categorisation, other studies have suggested that even young children are disposed to categorise objects according to presumed essences (Gelman & Medin, 1993; Gelman, 2000). Gelman & Wellman (1991) showed that 4 and 5 year old children appear to believe that an apple seed will grow into an apple tree, regardless of the environment in which this happens. Apparently children believe something inside the seed, and not contingent features of the environment, is causally responsible for the properties it later acquires.

In spite of the support essentialism has received, there have been counter claims. Malt (1994) showed that categorisation of instances of water is not fully explained by the proportion of H_2O people believe the instances contain. She argued for the importance of function instead. Braisby, Franks & Hampton (1996) showed that categorisation is at odds with predictions suggested by Putnam and Kripke's articulation of essentialism. Instead they argued categorisation was perspectival or context-sensitive. Kalish (1995) showed fuzziness in category boundaries that he argued was incompatible with essentialism. Yet, in a rejoinder, Diesendruck & Gelman (1999) have argued that findings such as these are compatible with essentialism.

This paper aims to add empirical evidence concerning psychological essentialism by examining an important implication of the essentialist view that has remained largely unexplored. Putnam (1975) developed a corollary of his essentialist view that he labelled the Division of Linguistic Labour. While being developed around word meaning, these arguments have nonetheless been taken to apply to concepts (e.g., Fodor, 1998). So interpreted they have the following implications. If categorisation is determined by micro-structural properties such as genetic, chemical or biological properties, then scientists who are expert in the appropriate domain are likely to have more information than lay-people on which to base their categorisations. If lay people are essentialist, they

should rationally defer to people with more knowledge of the relevant properties. For instance, if a metallurgist pronounces a gold watch to be "not gold," other things being equal, essentialism requires our categorizations to change accordingly. Deference arises from this division of linguistic labour – scientists are deemed to labour to uncover essential properties, while lay-people 'piggy-back' on their expertise. Putnam suggests there is a social dimension to concepts and categorization, one in which categorization by non-experts is intimately tied to, and parasitic on, categorization by experts.

There have been a number of recent theoretical examinations of deference (e.g., Fodor, 1998; though see Segal, 2000, for a different position), and different accounts proposed. Yet there has been no empirical evidence cited in support of these accounts. Similarly, studies have examined expertise in relation to categorization (e.g., Medin, Lynch, Coley & Atran, 1997), but have not been designed to tap the claims of essentialism. This paper seeks to establish, first, whether deference occurs and, second, parameters which govern that deference and, in so doing, offer a further evaluation of psychological essentialism.

The studies all examine the way in which genetically modified organisms are categorised. This is for two reasons. First, many prior examinations of essentialism have employed counterfactual scenarios involving fantastical transformations and/or discoveries of an object's properties. These scenarios may be hard to understand and use unfamiliar transformations. Second, few studies have examined how categorization changes as a function of changes in the information people possess about properties thought to be essential, such as genetic properties. Focusing on genetic modification allows the use of transformations of which people are likely to be aware, and allows a careful examination of the dependence of categorisation on genetic properties. It also allows the identification of groups thought to be expert and inexpert.

The first study examines the extent to which putative modifications in the genetic structure of organisms lead to changes in the way those organisms are categorized. Studies 2 and 3 examine the extent to which lay-people defer in categorisation. Study 2 examines deference to expert groups, predicted by essentialism, and study 3 functions as a control, examining deference to non-expert groups that is not predicted by essentialism.

Study 1

This study considers the way in which the categorisation of natural (living) kinds depends upon knowledge of the kind's genetic properties.

Design

Participants were randomly assigned to one of three conditions that varied according to the extent and nature of the modification – a Purification/Genetic Modification condition; a Same Super-ordinate category genetic modification condition; and an Other Super-ordinate category genetic modification condition.

Method

Participants 68 undergraduate psychology students attending an Open University residential school volunteered to participate.

Materials Four natural (living) kinds were chosen: apple, potato, salmon, chicken. These were chosen also to be food-stuffs so that they, and the prospect of their genetic modification, would be relatively familiar to the participants. Within these constraints, the kinds were also chosen to be as typical as possible of their immediate super-ordinate categories (i.e., fruit, vegetable, fish, bird).

Procedure Participants were presented with 8 scenarios, involving 2 different kinds of transformation for each natural kind category. In the Purification/Genetic Modification condition, half the scenarios referred to purification, half to genetic modification. In the Same Super-ordinate condition, transformations involved either 50% or approx. 100% of genetic material being taken from a member of the same super-ordinate category (e.g., for salmon, genetic material would come from other fish). In the Other Super-ordinate condition, transformations involved either 50% or approx. 100% of genetic material being taken from a category outside the super-ordinate (e.g., for salmon, from animals that are not fish).

The scenarios adopted the following form where X refers to one of the four kinds and Y refers to the relevant super-ordinate: "You have just bought an X from a reputable retailer. However, on examining its packaging closely, you discover that the X has been (genetically modified/purified, so as to remove many of the impurities often found in X/genetically modified, with around half of its genetic material coming from other Y/genetically modified, with nearly all of its genetic material coming from other Y/ genetically modified, with around half of its genetic material coming from [animals/plants] that are not Y/genetically modified, with nearly all of its genetic material coming from [animals/plants] that are not Y). In all other respects though the object looks, feels, smells and tastes just like an X." On reading each scenario, participants were asked to answer six questions, including a categorization question (Is the object that you have bought an X?).

Since the opening sentence of the scenarios refers to the object as a member of the category in question, this

procedure may lead to an underestimate of the impact of the transformations. However, this potential bias is difficult to avoid since failing to refer to the object as a member of a category would pragmatically imply that the object was thought not to be a member of the category, thus generating a potential opposing bias.

Results

Responses to the categorization question were analysed by a series of log-linear analyses (analysis of the other questions will not be reported). Different analyses were conducted for the three main conditions. The over-all results are shown in table 1.

Table 1. Percentage of Yes responses by condition

Transformation	% Yes responses
Purification	98.0
Modification	96.0
50% same super-ordinate	57.0
100% same super-ordinate	44.0
50% other super-ordinate	52.5
100% other super-ordinate	47.5

Purification/Genetic Modification Surprisingly, there was no effect of type of modification (i.e., purification vs. genetic modification), with 96.7% of all responses being Yes (i.e., responses that the purified or genetically modified object is a member of the kind).

Same Super-ordinate and Other Super-ordinate These conditions were combined for analysis. Categorisation depended upon the extent of the modification (i.e., 50% vs. approx. 100% genetic material being modified): when 50% of genetic material was modified, 55.0% of responses were Yes, which fell to 41.7% when approx. 100% of the material was modified (partial chi-square(1) = 12.45, $p < 0.001$). There was a marginal effect of the type of genetic material introduced: when material came from the same super-ordinate, 50.5% of responses were Yes, which fell to 45.6% when material came from another super-ordinate (partial chi-square(1) = 3.50, $p = 0.06$).

Discussion of Study 1

These results support the view that changes in genetic structure introduce changes in the way people categorize living kinds. While this is consistent with essentialism, what is striking about these results is how little categorization changes in the face of gross changes in genetic structure. Living kinds that have simply been 'genetically modified' are regarded almost universally as remaining members of the kind. Even when nearly all of a salmon's genetic material is said to come from animals that are not fish, approximately half of all responses still treat the object as a salmon. Given that humans and chimpanzees share approximately 98% of their DNA, the resistance of categorization to the influence of genetic modification is remarkable.

One explanation that is compatible with essentialism is that people's knowledge of genetic properties is so poor that the scenarios used here merely render them uncertain in their categorization. They may be unsure, for instance, whether genetic properties are likely to be essential or not. Indeed, a pattern of around 50% Yes and 50% no responses is suggestive of uncertainty. Another explanation, however, is that people are not only weighing the genetic properties of the objects, but also their appearance and functional properties and that, in these scenarios, they outweigh the genetic influence. Contra essentialism, this suggests that categorization is determined in part by non-micro-structural properties.

The patterns of approx. 50% Yes responses also imply that these scenarios are ideal for investigating Putnam's division of linguistic labour, since uncertainty is likely to increase the influence of expert opinion. This is the focus for studies 2 and 3.

Study 2

This study considers the way in which people's categorizations depend on those of expert scientists.

Design

Participants were randomly assigned to one of three conditions, selected from Study 1: a Genetic Modification condition; a 50% Same Super-ordinate condition; and a 50% Other Super-ordinate condition.

Method

Participants 90 undergraduate psychology students attending an Open University residential school volunteered to participate.

Materials The same materials as in Study 1 were used.

Procedure A similar procedure to Study 1 was used. However, information concerning how an expert group categorized each object was incorporated immediately after the description of the modification. In the genetic modification condition, the scenario read as follows: "You have just bought an X from a reputable retailer. However, on examining its packaging closely, you discover that the X has been genetically modified. According to most biologists the object (is/is not) an X. In all other respects though the object looks, feels, smells and tastes just like an X." Each of the 4 natural kinds was presented twice, once with an affirmative and once with a negative expert categorization judgement, yielding 8 scenarios. Participants were asked the same questions as in Study 1.

Results

Log-linear analyses were conducted for the three main conditions. Over-all results are shown in figure 1.

Genetic Modification Categorization varied according to how the biologists had judged the same categorizations: 87.5% of responses were Yes when the biologists had said Yes; 75.6% of responses were No when the biologists said No (partial chi-square(1) = 131.66, $p < 0.001$). Participants deferred more when the biologists said Yes than when the biologists said No (partial chi-square(1) = 6.57, $p < 0.05$).

50% Same Super-ordinate This condition yielded similar findings: 85.8% of categorization responses were Yes when the biologists had said Yes and 76.7% of responses were No when the biologists said No (partial chi-square(1) = 125.93, $p < 0.001$). Participants again deferred more when the biologists said Yes (partial chi-square(1) = 3.86, $p < 0.05$).

50% Other Super-ordinate Similar findings emerged: 62.9% of responses were Yes when the biologists had said Yes and 72.7% of responses were No when the biologists said No (partial chi-square(1) = 32.94, $p < 0.001$). This time, however, participants deferred more when the biologists said No than when the biologists said Yes (partial chi-square(1) = 3.32, $p < 0.05$).

Further analysis showed the number of Yes responses differed across these two latter conditions (partial chi-square(1) = 6.85, $p < 0.01$). Also, the dependence of categorization on the Biologists' prior categorization differed across these two conditions (partial chi-square(1) = 7.58, $p < 0.01$). These results are discussed in conjunction with those of Study 3.

Study 3

This study considers the way in which people's categorizations depend on those of non-experts.

Design

Participants were assigned to conditions as in Study 2.

Method

Participants 62 psychology students attending an Open University residential school volunteered to participate.

Materials The same materials as in Study 1 were used.

Procedure A similar procedure to Study 2 except that information concerning an expert group's categorization was replaced by information about a non-expert group's categorization. The word "biologists" was replaced with the word "shoppers" to produce the scenarios. Again, each natural kind was presented twice, once with an affirmative and once with a negative non-expert judgement. Participants were asked the same questions as in Studies 1 and 2.

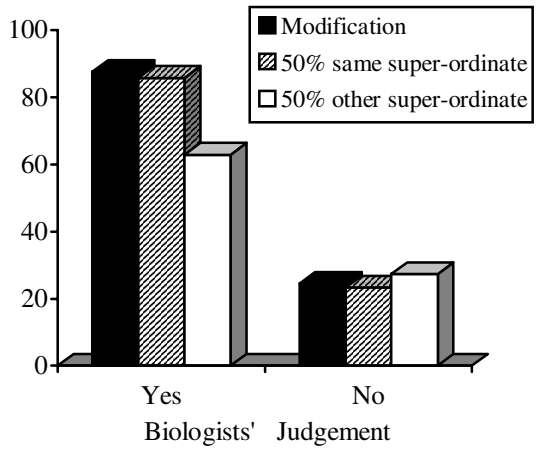


Figure 1. Percentage Yes responses by transformation type and the biologists' judgements

Results

Similar analyses to study 2 were conducted. Over-all results are shown in figure 2.

Genetic Modification Categorization varied according to how the shoppers had judged the same categorizations: 86.4.% of responses were Yes when the shoppers had said Yes; 36.4% of responses were No when the shoppers said No (partial chi-square(1) = 18.92, $p < 0.001$). Participants deferred more when the shoppers said Yes than when the shoppers said No (partial chi-square(1) = 53.24, $p < 0.001$).

50% Same Super-ordinate This condition yielded similar results: 71.3% of responses were Yes when the shoppers said Yes; 55.0% of responses were No when the shoppers said No (partial chi-square(1) = 19.02, $p < 0.001$). Deference was greater when the shoppers said Yes (partial chi-square(1) = 4.81, $p < 0.05$).

50% Other Super-ordinate Rather different results emerged: 44.3% of responses were Yes when the shoppers said Yes; 68.8% of responses were No when the shoppers said No (partial chi-square(1) = 4.16, $p < 0.05$). This deference was greater when the shoppers said No (partial chi-square(1) = 10.30, $p < 0.01$).

Further analysis showed that the number of Yes responses varied across these two latter conditions (partial chi-square(1) = 15.81, $p < 0.001$). Also, the dependence of categorization on the Shoppers' categorization varied marginally by these two conditions (partial chi-square(1) = 3.66, $p = 0.06$).

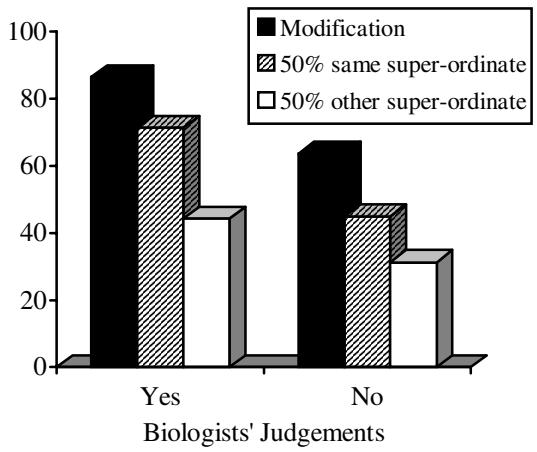


Figure 2. Percentage Yes responses by transformation type and by the shoppers' judgements

Discussion of Studies 2 and 3

Studies 2 and 3 show a complex pattern of responding. In study 2, most responses made by participants conform to those of the expert group (biologists). Superficially at least, this appears to show support for essentialism, and Putnam's division of linguistic labour. Nonetheless, study 2 also shows other influences on categorization, ones that are not so easily explained by an essentialist account. Firstly, there is a substantial minority of responses that are in opposition to those of the biologists. For the 50% Other Super-ordinate condition, there is greater opposition when the biologists' judge the organism to be a member of the kind. For the 50% Same Super-ordinate condition, there is greater opposition when the biologists' judge the organism not to be a member of the kind. While it would be difficult to argue these completely undermine the apparent support for essentialism that the majority of responses show, they do raise the question as to why participants are choosing to categorise in opposition to the experts, and so presumably in opposition to the micro-structural properties. They also raise the question as to how much deference essentialism predicts.

The findings of study 2 suggest that while people do defer to experts, they do so less when the experts' judgements are contrary to their own – the deference is conditional. The question then arises as to what is influencing people's categorization so strongly that they will disregard the opinions of experts to whom they will defer in other circumstances. One possibility is that these findings point towards a continuing role for appearance and functional properties in categorizing natural kinds.

Participants in all of the studies were given the opportunity to offer written comments, and some of these support the importance of appearance and functional properties. One participant wrote: "If biologists said it was a fruit, that's OK. If biologists said it was not a fruit, that's not OK." This comment bolsters the view that the pattern of deference that is revealed is conditional. Another participant wrote "The apple we eat is not defined by biologists but by how it looks and tastes to non-experts" thus revealing a strongly non-deferential position.

These suggestions from study 2 are supported by study 3. Since shoppers generally are not expert with regard to the genetic properties of organisms, there is no essentialist basis for people to defer to this group. Nonetheless, this study revealed deference, albeit to a much smaller degree than in study 2. It also revealed similar systematic variation in deference. For the 50% Same Super-ordinate condition, people showed greater conformity when the shoppers judged the organism to be a member of the kind. For the 50% Other Super-ordinate condition, there was greater conformity with the shoppers' judgement when the shoppers judged the organism not to be a member of the kind. Again, it may be that people are more willing to 'defer' when the shoppers' judgement conforms to their own.

General Discussion

Overall, it seems as if deference to expert groups occurs, and more so than to non-expert groups. However, the extent of the deference depends upon how the organism has been transformed, and on what categorization judgement the expert group gives. This then is a pattern of partial or conditional deference.

Deference to non-expert groups also occurs and this raises questions concerning the basis on which people may defer in categorization. If people defer to others, such as shoppers, who do not possess expertise with regard to the relevant micro-structural properties, then we may also question why people defer to experts. Could it be that people defer to experts not because of their presumed greater knowledge of micro-structural properties?

The systematic influences on deference, coupled with the striking resistance of categorization to gross changes in genetic properties, suggest that categorization is influenced both by micro-structural properties and by appearance and functional properties.

This may be explicable on a perspectival view of concepts, according to which concepts have multiple contents that shift systematically according to perspective and context (Braisby, 1998). On such a view, concepts might reflect essential, micro-structural properties from some perspectives, but appearance and/or functional properties from others. These findings would then reveal a conflict for people seeking to categorize natural kinds: between deferring to experts on the micro-structural properties on the one hand, while being influenced by appearance and functional properties on the other. If this is right, then the findings

reported here suggest that essentialism can provide only a partial explanation of concepts and categorization.

Nonetheless, these studies represent just the first step in a wider programme of much needed research, one that raises many difficult questions. These studies have used objects whose appearance and functional properties are stipulated to be fixed, and whose micro-structural properties are manipulated. What would happen under the reverse conditions? Essentialism would predict little impact of changes in appearance and function relative to micro-structural properties. How should deference be operationalised? It has been operationalised in these studies as a switch in categorization due to the categorizations of others. But are there other, better ways of operationalising it? Finally, how are we to make sense of the interplay between categorization and social influences? Finally, what is the relation between deference and conformity or compliance? These studies suggest a fruitful interaction between social psychological work on these issues and the cognitive psychology of categorization.

Acknowledgments

I would like to thank Bradley Franks for discussions of the ideas contained herein; any errors remain my own.

References

Braisby, N. (1998). Compositionality and the modelling of complex concepts. *Minds and Machines, 8(4)*, 479-508.

Braisby, N., Franks, B., & Hampton, J. (1996). *Essentialism, word use, and concepts. Cognition, 59*, 247-274.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge: MIT Press.

Diesendruck G. & Gelman S. A.(1999). Domain differences in absolute judgments of category membership: Evidence for an essentialist account of categorization. *Psychonomic Bulletin & Review, 6(2)*, 338-346.

Fodor, J. (1998). *Concepts: Where cognitive science went wrong*. Oxford: Oxford University Press.

Gelman, S. A. (2000). The role of essentialism in children's concepts. *Advances in Child Development and Behavior, 27*, 55-98.

Gelman, S. A. & Medin, D. L. (1993). What's so essential about essentialism? A different perspective on the interaction of perception, language, and conceptual knowledge. *Cognitive Development, 8*, 157-167.

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the nonobvious. *Cognition, 38*, 213-244.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman, *Problems and projects.* Indianapolis, IN.: Bobbs-Merrill.

Kalish C. W. (1995). Essentialism and graded membership in animal and artifact categories. *Memory & Cognition, 23(3)*, 335-353.

Kaplan, A. S. & Murphy, G. L. (2000) Category learning with minimal prior knowledge. *Journal of Experimental Psychology: Learning Memory And Cognition, 26(4)*, 829-846.

Keil, F. (1986). Conceptual development and category structure. In U. Neisser (Ed.), *Concepts and conceptual development*. Cambridge: Cambridge University Press.

Keil, F. C. (1989). *Concepts, kinds and cognitive development*. Cambridge: MIT Press.

Kripke, S. (1980). *Naming and necessity*. Cambridge: Harvard University Press.

Malt, B. C. (1994). Water is not H_2O. *Cognitive Psychology, 27*, 41-70.

Margolis, E. & Laurence, S. (1999). Introduction. In E. Margolis & S. Laurence (Eds.), *Concepts: Core readings*. Cambridge, MA.: MIT Press.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist, 44*, 1469-1481.

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology, 32*, 49-96.

Medin, D. L. & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou and A. Ortony (Eds.), *Similarity and analogical reasoning*. New York: Cambridge University Press.

Murphy, G. L. (1993). Theories and concept formation. In I. Van Mechelen, J. Hampton, R. Michalski & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis*. London: Academic Press.

Murphy, G. L. & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289-316.

Putnam, H. (1975). The meaning of 'meaning.' In H. Putnam, *Mind, language, and reality: Philosophical papers, vol. 2*. Cambridge: Cambridge University Press.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. New York: Cambridge University Press.

Rosch, E. H. (1975).Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*, 192-233.

Rosch, E. H. & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573-605.

Segal, G. (2000). *A slim book about narrow content*. Cambridge, MA.: MIT Press.

Spalding, T. L. & Murphy G. L. (1996). Effects of background knowledge on category construction. *Journal of Experimental Psychology: Learning Memory and Cognition, 22(2)*, 525-538.

Meaning, Communication and Theory of Mind.

Richard Breheny (reb35@cam.ac.uk)
RCEAL, University of Cambridge, Trumpington Street
Cambridge, CB2 1QA UK

Abstract

The study of language, meaning and communication in the cognitive sciences has undergone a kind of conceptual inflation in the past twenty years or so. Not only has the very nature of human communication come to be seen as, in many respects, Gricean, but also linguistic meaning itself has come to be widely regarded in terms of the effect of language use on mental states. As a result, a more or less explicit assumption about the conceptual abilities of agents who have linguistic and communicative competence has been adopted in a variety of disciplines ranging from language acquisition to formal semantic theories: that these agents have the ability to represent and make inferences about the mental states of others. The purpose of this paper will be to offer considerations in support of the contrary, more minimalist view that neither meaning nor communication involve the representation of mental states essentially. Correspondingly, agents who are competent with regards language use and communication need not possess meta-cognitive abilities.

Introduction

The study of language, meaning and communication in the cognitive sciences has undergone a kind of conceptual inflation in the past twenty years or so. Not only has the very nature of human communication come to be seen as, in many respects, Gricean, but also linguistic meaning itself has come to be widely regarded in terms of the effect of language use on mental states. As a result, a more or less explicit assumption about the conceptual abilities of agents who have linguistic and communicative competence has been adopted in a variety of disciplines ranging from language acquisition to formal semantic theories: that these agents have the ability to represent and make inferences about the mental states of others. The purpose of this paper will be to offer considerations in support of the contrary, more minimalist view that neither meaning nor communication involve the representation of mental states essentially. Correspondingly, agents who are competent with regards language use and communication need not possess meta-cognitive abilities.

The Dilemma.

Different theories of language and communication presuppose different kinds of cognitive capacities - either explicitly or implicitly. Among the more prominent and most influential pragmatic theories - theories of speech acts, conversational implicature and the like - are theories which are broadly Gricean in their stance. Gricean theories can be defined as those theories which analyse utterances as acts by one agent which seek to alter the mental states or attitudes of other agents in part by getting the other agent to recognise their intention to so do. It follows that Gricean approaches to pragmatics presume that communicating agents posses the cognitive ability to represent the mental state or attitudes of other agents and/or to make inferences about these.

Of course, Grice's theory of conversation as presented in his "Logic and Conversation" (Grice 1975) contains a working-out schema for conversational implicature which is a piece of pure belief-desire psychology, with inferences being explicitly made about the attitudes of another agent. At a perhaps more fundamental level, influential theories of basic speech acts such as assertions adopt a more or less Gricean stance. Stalnaker's speaker presupposition framework, in particular, presumes that agents involved in conversation assume a common ground. A proposition is common ground, or presupposed by the speaker, if the speaker is disposed to act as if she believes it or assumes it is true and believes that her audience believes or assumes it is true. Assertions and suppositions are acts which seek "to change the presuppositions of the participants in the conversation by adding what is asserted to what is presupposed". (Stalnaker 1978:323). Thus, according to Stalnaker's model of assertion, in order to engage in conversation one must be able to represent speaker presupposition. And the structure of this presupposition "can be represented by a Kripke semantics in which the accessibility relation is serial, transitive and Euclidean, but not necessarily reflexive" (Stalnaker 1996:282). In other words, putting aside certain idiosyncrasies, speaker presupposition is structurally similar to other attitudes and therefore requires similar conceptual abilities to represent it. Other influential accounts of speech acts fundamentally incorporate some notion of common ground with basically the same structure - see Searle (1969), Lewis (1969), Schiffer (1972), and more recently Clark (1996). Sperber and Wilson's Relevance Theory (1986/95) also supposes that basic assertive

speech acts involve the recognition of complex intentions involving the intention to get the audience to believe what the speaker is saying.

These Gricean pragmatic theories have also inspired an approach to meaning which has been popular in the recent past. Consider again Stalnaker's proposal regarding assertions. They are seen as moves which are made on the common ground, a proposal to reduce the set of live possibilities consistent with what is presupposed in accordance with the content of what is said. In this framework, the meaning of the linguistic expressions used was thought about in traditional truth-conditional terms. Dynamic semantics (Kamp 1981, Heim 1982, Groenendijk & Stokhof 1990, 1991) takes the further step of supposing that the meaning of a sentence consists in its potential for transforming the input context set into the resultant output state. Thus meaning of sub-sentential elements lies in their contribution to the update potential of the sentences. So we could say that in dynamic approaches, the meaning of a predicate like "sleeps" no longer makes reference just to the property of sleeping or some such notion which would be central in stating the predicate's contribution to truth-conditions (say, a function from individuals to truth-values), but the predicate's meaning also involves an input state - something which has the same structure as Stalnaker's speaker presupposition. That is, it would be a function from individuals to a function from input states to output states. Thus dynamic semantics imputes to language users who can be said to know or grasp the meanings of basic expressions in their language, this sophisticated ability to represent mental states.

In summary, both at the level of semantic and pragmatic theory, it is a widely held assumption that agents who engage in basic communication are capable of thinking about or representing other agents as bearing propositional attitude-type relations to what is being communicated. However, it is also a widely held assumption in psychology that children under the age of four years do not posses this ability. This assumption is founded on a fairly impressive and largely conclusive body of experimental work over the past decade or so, starting with Wimmer & Perner (1983). So there is a tension between what these influential semantic and pragmatic theories assume about language users in general and what experimental evidence suggests about a significant minority of them. In the balance of this paper, we will consider three options for relieving this tension. Option I: We could argue that young children do not ever properly engage in communication and (optionally), that young children do not really understand the meaning of the expressions they use. Option II: We could challenge the results concerning so-called theory of mind abilities in young children. Option III: We could say that the above assayed theories do not capture the essence of communication but, at best, only the norm among sophisticated language users who have theory of mind abilities.

Are young children competent communicators?

The viability of Option I depends on how easily one can overturn the prima facie intuition that young children, aged two to three years, are capable communicators in the following sense: in at least some cases, their use of language or their understanding of others' use of language is at a level of performance equivalent to that of an adult. That is, in at least some situations when a child utters a sentence, S, their intentions with regards the content of the utterance are clearly comprehensible and are the same as those a normal adult would be attributed with if it uttered S in the same circumstances. Similarly, in at least some cases where a child is faced with an utterance of S by another agent, their grasp of that action is the same as that of an adult faced with the same utterance.

Of course, we agree that children of this age are not nearly as good at communication as adults. They are much more prone than adults to misunderstanding, mis-communications, irrelevancies and so on. Also their linguistic proficiency is in many ways not the same as adults. In particular they have a much more limited vocabulary. But this is a matter of degree. They do have the basic wherewithal to engage in linguistic communication, in spite of the fact that their cognitive capacities limit the degree of success in this matter.

For us to take Option I seriously, we would need a lot more evidence that children are not competent when it comes to basic communication. Presently, it does not seem all that likely that this evidence would be forthcoming. Consider for instance personal pronouns ("she", "he", "it" etc). These are among the first words children learn (Bloom 2000). Moreover, their usage of these forms evinces a more or less adult competence in circumstances where there are no extra demands placed on the child which are beyond their conceptual abilities. This particular fact is significant, given that Gricean-Stalnakerian theories of pronoun usage by and large attach sophisticated presuppositions (involving the common ground) to pronouns.

In the absence of any strong arguments for this option, we will put it aside and move on to consider the other two.

Challenging the theory of mind orthodoxy.

Option II seems far more promising in the light of recent research into word learning. Here the suggestion is that children younger than four years old have a much more sophisticated appreciation of others' mental

states than the classic Sally-Anne experiments suggest. There are two important strands to this argument which we need to consider here. Both are raised in Bloom & German (2000).

The first line of attack would be to question the assumption that the Sally-Anne task probes the onset of full theory of mind abilities. Bloom & German argue that this kind of false belief task involves abilities other than theory of mind (ibid:B26). In particular, they claim, citing a variety of experimental evidence, that it is reasoning about false beliefs that causes difficulty for children who otherwise might reasonably be supposed to have theory of mind ability. That is, false-belief tasks are difficult for young children because of the difficulties generally attached to reasoning about falsehoods rather than because they lack theory of mind abilities.

Experiments which are designed to lighten subjects' processing load have been found to facilitate performance. For example, German & Leslie's (2000) modified false belief tasks lowered the passing age by a few months. These results could be seen as significant in the context of theories which suggest that theory of mind abilities are in some sense modular. In the tradition of modular approaches to the mind, one could argue that young children's theory of mind module is 'switched on' or 'matures' earlier than classical Sally-Anne tasks suggest, but that due to the processing load demanded by reasoning about false beliefs, children fail.

Bloom & German argue that results from other experiments provides support for this view. These experiments involve thinking about non-actual states of the world but do not involve folk-psychological reasoning as such. The 'false photograph task' has the same structure as the false belief task except that it does not involve thinking about mental states. That is, children are asked about the content of a photograph when it does not match the current state of the world. Three year old children who fail false-belief tasks also fail the false photograph task (Leslie 2000). Other related evidence mentioned by Bloom & German involves children's performance on tasks involving counterfactuals. Their conclusion is that it is not necessary that children fail false-beliefs tasks because they do not have a working theory of mind. Moreover they suggest that it is more the general difficulty of the task which bars success. Bloom and German go on to cite positive evidence for younger children's theory of mind ability. Before we consider this important evidence, let us consider this first line of attack: Children fail false belief tasks because certain elements of the task are beyond them. These elements arise in non-theory of mind tasks such as the false photograph task and tasks involving counterfactuals so it is not lack of theory of mind abilities which is responsible. If this line of argument seems appealing at first, a moment's thought should reveal that it has things the wrong way around.

The false belief task was originally designed on reflection about the nature of theory of mind. Having a theory of mind means (at least) having an ability to think about the actions of other agents as governed by causally active, but unobservable, mental states. This ability presupposes having an ability to represent an agent as having propositional attitudes. Even if another agent has a true belief, representing that fact requires conceptual abilities far different from representing the content of that belief. The conceptual abilities involve an appreciation of the different accessibility relations that need to be associated with different agents. That is to say, according to one popular metaphor, one needs to set up different belief boxes (and desire boxes etc) for different agents.

One could argue that certain cognitive and conceptual abilities required for the false photograph task, for tasks involving counterfactual states and others are the same as those required for theory of mind tasks. In particular, there is a strong case to be made for the claim that to perform these latter tasks, one needs to think with different frames, using different accessibility relations. What this means in cognitive terms is something of an open question. At a minimum, it means over-riding basic dispositions regarding the representation of two situations. Consider, for instance, the false photograph task (Zaitchik 1990). The subject sees a Polaroid photo being taken of a scene in which a cat is on the mat. As the photo is developing, the subject sees the experimenter change things in the scene so that the cat is no longer on the mat. The child is asked, "In the photograph, where is the cat sitting?". In order to successfully complete the task, the child has to represent the situation in the photo, s', as well as the current situation, s. Now, normally if the child represents s and s' then it can infer that there is a situation, s'', which contains both. It would also be disposed to reject (or suppress) representations of one of two incompatible situations. To perform the task, these basic dispositions have to be overridden. It does not seem plausible that such basic inferences or processes would be overridden except where there are two different frames under consideration. That is, why else would the cognitive system develop a mechanism whereby these fundamental dispositions are forestalled?

So, contrary to Bloom & German, we should conclude from these experiments that there is no evidence that three year-olds posses the kind of abilities which are pre-requisite for having theory of mind.

Bloom & German's second line of argument has more substance. It is based on a growing body of experimental work in word learning and other developmental research which is at least as impressive

as the false-belief literature. I will mention briefly some key results here before discussing the third alternative. In the light of that discussion, I will propose that what may seem to be evidence of genuine theory of mind ability could equally well be accounted for in terms of an independently motivated ability of children to keep track of an object of joint attention between themselves and other agents. This ability does not presuppose those required for theory of mind tasks.

The crucial data for precocious theory of mind abilities comes from investigations which seek to establish the role in word learning of the interactional dimension of communication (joint attention etc) and children's appreciation of other agents as intentional - what Tomasello calls 'social cognition'. The data reviewed in Tomasello (1995), Tomasello (2000), Bloom (2000) involves experiments where young children (2-3 years) display an appreciation of others' intentions and apparently of others' mental states (ignorance) when learning words. For instance, Tomasello and Barton (1994) discuss an experiment where an adult announces that it is going to find a toma (a novel word) standing over a number of opaque containers. From each, the adult produces novel objects and reacts in a disappointed fashion to all but one to which she responds in a manner appropriate to successful finding. Afterwards, the child subject is tested to see whether it has learned the word 'toma'. The results are that the subjects learn the word as applying to the 'found' object, suggesting that the children are sensitive to the adult's intentions in such situations.

More interestingly, in their communicative behaviour, children seem to show an appreciation of adults' ignorance in both word learning scenarios (Akhtar, Carpenter & Tomasello 1996) and other scenarios (O'Neill 1996). In the former case, a child learns a word when it is mentioned by a parent in a context where there is one novel item for the parent and three other items which the parent and child had just played with but which had remained unnamed. In matching the novel word to the novel item, the children seem to be displaying an appreciation of the mental states of the parent. Perhaps even more interestingly, Happe & Loth (in press) have results from a word learning task based on the structure of Sally-Anne which suggests that children who fail the false-belief task manage to learn a word under the same conditions. Ie. Sally and child subject play with novel toy. Sally puts toy in container A and goes away. Anne comes with her own novel object. Anne makes the switch with her toy and Sally's. Sally returns and says, pointing to where she left her object, "Let's play with the modi". Children who fail basic false-belief tasks perform better at learning 'modi' as applying to not what is in the box but what Sally had put in there. What, then, can explain this apparent sensitivity on the part of children to the intentions and mental states of other agents, other than theory of mind? The answer to this question does not involve any kind of mysterious interim ability on the part of children. It can be found by thinking carefully through a developmental path commensurate with children's developing communicative and linguistic abilities.

Basic communication.

Although Sperber & Wilson are somewhat culpable in this conceptual inflation when it comes to communication, the essence of their theory is built around a much more parsimonious view: an act of communication is simply an act whereby one agent attempts to draw another agent's attention to something. They contend that agents to whom this kind of behaviour is directed decide on what their attention is being directed to by processing input stimuli for relevance - which is defined in terms of a kind of cognitive nutrition and processing effort. The food metaphor is apposite when we consider how a pre-linguistic child might come to respond to ostensive behaviour in this way and to eventually produce such behaviours itself.

The key to communicative development comes with conceptual development around 9-12 months. This development (surveyed in Tomasello 1995) involves the formation of concepts of actions. At this stage a child begins co-ordinating first person experience with memories of observed behaviour of others, with kinaesthetic memories etc. It is implicit in this abstraction over experience that there is an agent of the action (among other participants) and there are constituent acts in the action. Also, an action concept would be associated with episodic memory of prototypical situation types in which the action takes place. These in turn have constituent eventualities, including typical end states. Eating, for example, will be conceptualised as consisting of certain actions, and it will be associated with certain typical types of situation, including the end state associated with tasting and swallowing the food. Forming concepts of actions which are directed toward another agent presumably does not involve any extra conceptual abilities. Feeding, for a typical example, would be conceptualised in terms of constituent acts on the part of one agent directed towards another.

In general, recognising an action A as such does not presuppose any special abilities beyond this ability to break it down into constituent acts and to keep track of information about the typical eventualities involved. In particular, it does not involve theory of mind. Also, it does not presuppose that one witness the whole act, just some constituents of the action would normally suffice to trigger recognition. Hence a child can recognise a

failed attempt at A as such. This fact could be used to account for the data in Barton & Tomasello (1994).

Looking (or attending to) is an act which we can suppose that children with these basic abilities can conceptualise. It is an act directed toward a situation (in the sense of a chunk of the world as per Barwise and Perry 1983) which results (potentially) in certain cognitively nutritional effects. Contrary to Tomasello (1995) and others, joint attention need not involve mutual knowledge or any special social-cognitive skills. It is just a matter of following into the gaze of another (presumably in the hope of cognitive effects). Gaze monitoring is just a matter of monitoring the actions of another (again, possibly for reasons of self-interest). Showing and other ostensive acts, like feeding, are just actions on another. The third participant role in this kind of act is not filled by food but a situation. As mentioned above, as with gaze monitoring, children would naturally process such acts for relevance. So like feeding, it is a benevolent act. Why it is that children themselves come to show things to others is not clear - but nor is it clear why they offer food or engage in other reciprocating benevolent behaviours. That children do offer up things for attention would explain their inclination to indicate new things to their parents and other carers. It would provide for an alternative account of Akhtar, Carpenter & Tomasello's (1996) finding, if we assume the child can keep track of what objects the relevance of which it has and has not shared with significant others. We will see briefly that there is good independent evidence for this. In that case, when the adult returns in Akhtar et al's crucial condition, the child would be focussing on the newest toy since that is something yet to be shared. Thus the child will assume that the adult's (purposely vague) indicating will be directed toward the new toy since that will be its first accessible interpretation. (See Sperber 1994 for a discussion of different relevance-based interpretation strategies for individuals with different levels of theory of mind abilities).

With joint attention and with directing attention, there is not only a situational object but a larger situation involving the two agents. With Tomasello, we agree that language is acquired in the context of such interactions. Words and sentences are constituents of ostensive acts, being descriptive of the type of situation being indicated. Pronouns are learnt as acts of pointing at objects in the situation being indicated. There need be no Stalnakerian presupposition for the proper mastery of these forms (although sophisticated adults can optionally make such presuppositions - to the effect that the referent of a pronoun is in the object of joint attention). Thus children, like anyone else, can engage in communicative activities without concerning themselves with speaker presuppositions. Indeed, where young children need to take common ground into account to succeed, they tend to fail. For instance, Mitchell et al (1999) devised a Sally-Anne task with referring expressions (descriptions) and the results were predictably that three year-olds failed and four year-olds passed. So what is the difference between this and Happe and Loth's word learning case? Crucially, in the latter, it can be argued that the child can complete the task successfully simply by being able to track what the object of joint attention is between itself and a number of agents. In Mitchell et al's task, as with Sally-Anne, success depends on thinking about the mental state (common ground) of another agent. Notably, in Happe and Loth's study, they did a so-called 'true belief' version of the word learning task. On this task, children under four performed worse than in the corresponding true-belief Sally-Anne task but they performed with the same level of success as with the false-belief word-learning task. Happe and Loth have no explanation for this but there is an explanation given the focus of attention account: The 'true-belief' word-learning task involves exactly the same skills and demands as the 'false-belief' word-learning task. Sally puts object X in A and along comes Anne and introduces object Y. She replaces X with Y in A in Sally's presence. When Sally stands over A and says, "Let's play with the modi", there is understandable confusion since the child has presumably been tracking Anne's gaze on Y and not Sally's.

With other cases where children seem to be sensitive to what other agents do and do not know, proper attention to their abilities to track the objects of joint attention and their relevance guided abilities to lock onto what is being indicated would reveal that they are not so sophisticated after all.

Conclusion.

It seems fairly clear-cut what theory of mind is and what conceptual abilities it entails. The dominant tradition in pragmatics and the dynamic tradition in semantics presumes that language users have theory of mind or, at least, the conceptual abilities which underpin theory of mind. Young children do not possess these abilities and yet they seem to communicate perfectly adequately and they seem to have a firm grasp on the meaning of at least some basic expressions in their language. If we accept this, then we have to say that Gricean ideas about language use only apply to more sophisticated language users. A more minimal theory of basic communication has been offered here based around some ideas from situation theory and relevance theory. To be sure, according to the alternative suggested here, no communicative abilities can get off the ground without a child having certain affinities with other agents. In particular, the development of concepts of actions clearly entails co-ordinating first person and third person experience.

However, we have suggested here a way of thinking through social development which does not call for any mysterious interim psychological appreciation.

References

Akhtar, N., M. Carpenter & M. Tomasello (1996). The role of discourse novelty in children's early word learning. Child Development. 67: 635-645.

Barwise, J. & J. Perry (1983). Situations and Attitudes. Cambridge Mass.: MIT Press.

Bloom, P. (2000). How Children Learn the Meanings of Words. Cambridge Ma. MIT Press.

Bloom, P. & T.P. German (2000). Two reasons to abandon the false belief task as a test of theory of mind. Cognition 77: B25-B31.

Clark, H. (1996). Using Language. Cambridge, CUP.

German, T.P. & A. Leslie (2000). Attending to and learning about mental states. In P. Mitchell & K. Riggs (eds) Children's Reasoning and the Mind. Hove: Psychology Press.

Grice, H.P. 1975. Logic and Conversation, in Syntax and Semantics 3: Speech Acts. (eds) P. Cole and J. Morgan, Academic Press, NY. pages 41 - 58.

Groenendijk, J. & M. Stokhof (1990). Dynamic Montague Grammar. In Kálmán L. & L. Pólos (eds.) Papers from the Second Symposium on Logic and Language. Budapest: Akadémiai Kiadó.

Groenendijk, J. & M. Stokhof (1991). Dynamic predicate logic. In Linguistics & Philosophy 14: 39-100.

Happe, F. & E. Loth (ms) Words speak louder than actions: children track false beliefs to learn new words before they can pass 'false belief tasks'. (to appear in Cognition)

Heim I. 1982. The Semantics of Definite and Indefinite Noun Phrases. PhD diss. U. Mass. Amherst.

Kamp, H. (1981). A Theory of Truth and Semantic Representation. In J. Groenendijk et al. (eds.) Truth, Interpretation and Information. Dordrecht: Foris.

Leslie, A. (2000). How to acquire a 'representational theory of mind'. In D. Sperber & S. Davies (eds) Metarepresentation. Oxford: OUP.

Lewis, D. (1969). Convention. Cambridge, Ma: Harvard University Press.

Mitchell, P., E J. Robinson & D E. Thompson (1999) Children's understanding that utterances emanate from minds: using speaker belief to aid interpretation. Cognition. 72: 45-66

O Neill, D K. (1996). Two year-old children's sensitivity to parent's knowledge state when making requests. Child Development. 67: 659-677.

Schiffer, S. (1972). Meaning. Oxford: Clarendon Press.

Searle, J. (1969). Speech Acts. Cambridge: CUP.

Sperber, D. (1994). Understanding verbal understanding. In Jean Khalfa (ed.) What is Intelligence? Cambridge, Cambridge University Press. 179-198.

Sperber D. & D. Wilson (1986). Relevance: communication and cognition. Oxford: Blackwell. (2nd edition 1995).

Stalnaker, R. (1979). Assertion. In P. Cole (ed.) Syntax and Semantics vol. 9: Pragmatics. New York: Academic Press.

Stalnaker, R. (1996). On the representation of context. In T. Galloway & J. Spence (eds.), Proceedings of SALT VI. Cornell University Press. pp279-294

Tomasello, M. (1995). Joint attention as social cognition. In C. Moore & P. Dunham (eds), Joint Attention: Its Origins and Role in Development. Hillsdale, NJ: Lawrence Erlbaum.

Tomasello, M. (2000) Perceiving intentions and learning words in the second year of life. In M. Bowerman & S. Levinson (eds) Language Acquisition and Conceptual Development. Cambridge: CUP.

Tomasello, M. & Barton, M. (1994). Learning words in non-ostensive contexts. Cognitive Development. 10: 201-224.

Wimmer, H. & J. Perner (1983). Beliefs about beliefs: representation and the containing function of wrong beliefs in young children's understanding of deception. Cognition 13: 103-128.

The Effects of Reducing Information on a Modified Prisoner's Dilemma Game

Jay C. Brown (Jaybrown@andrew.cmu.edu)
Carnegie Mellon University
Department of Psychology
Pittsburgh, PA 15213

Marsha C. Lovett (Lovett@andrew.cmu.edu)
Carnegie Mellon University
Department of Psychology
Pittsburgh, PA 15213

Abstract

Participants played a modified prisoner's dilemma game in which competition was created using a single player. The competition was between the player at the moment and the player in the future. The complexity of the game was increased across experiments. The transition from Experiment 1a to 1b saw the removal of information about future consequences and past behavior. Experiment 2 removed information about current outcomes. As the complexity of the game increased (both quantitatively and qualitatively) and therefore the external validity increased, the ability to "solve" the game decreased.

Introduction

Many choices we make are between things that make us feel good at the moment and things that are actually better for us in the long run. Choosing to consume alcohol at a party certainly feels better (at the moment) than not consuming alcohol, but in the long run (hangovers, loss of peer respect, etc.) we are certainly better to refrain from this consumption. When purchasing an automobile, assuming equal price, a sports car is definitely flashier than a mini-van, but the mini-van will probably last longer, be more practical, and cost less in insurance. Purchasing the sports car may make us feel better at the moment, but the total utility (over the life of the vehicles) would unquestionably be higher for the mini-van.

Impulsiveness is defined as the choice for the outcome that feels good at the moment (consuming alcohol; the sports car). Self-control is the choice for the outcome that is actually better in the long-run (refraining from alcohol consumption; the mini-van). Many factors affect our impulsiveness and ability to exhibit self-control. The experiments presented here address several of them.

A goal of this paper is to explore how people learn to choose between impulsivity and self-control. One can view this learning as an adjustment of strategy choices after feedback. Given the task studied in this paper, another relevant perspective views the learning as a growing understanding of cooperation in an iterated prisoners dilemma game.

In a traditional prisoner's dilemma game, two players each choose between two options (often called cooperate and defect) creating four possible outcomes (Rapoport & Chammah, 1965). These outcomes are associated with different rewards, labeled A, B, C and D, that must obey the following rules:

$$B > A > D > C$$
$$2A > B + C > 2D$$

Both players will receive outcome A (moderately good) if both choose to cooperate. Both players will receive outcome D (moderately bad) if both choose to defect. However, if Player 1 chooses to cooperate and Player 2 chooses to defect, then Player 1 will receive outcome C (the worst) while Player 2 will receive outcome B (the best). Defection tends to dominate in both one-shot and iterated playing of this game. However, if the players know each other, and, more importantly, trust each other, then cooperation can arise and persist. Rachlin, Brown and Baker (2001) have shown that Player 1 will cooperate only if he or she believes Player 2 will reciprocate that cooperation.

This result suggests that converting the two-player version of the game to a single-player version[1] would lead to high levels of cooperation. Nevertheless, previous work has shown that individual players chose to "cooperate" with themselves only 54% of the time (Brown & Rachlin, 1999). Why is this percentage so low? Consider the competition engendered by the single-player game: it is between the self at the moment and the self in the future, which is essentially a choice between impulsivity and self-control. The present experiments explore the processes by which individuals choose between these options in the single-player game.

Experiment 1a

The first experiment was designed to be a computer-based replication of previous work (see Brown & Rachlin, 1999 for full details) to obtain greater control

[1] A single-player prisoner's dilemma game, where that player makes two or more sequential choices, is identical to a two-player game in which the "other" player uses the tit-for-tat strategy perfectly (see Axelrod, 1987). In both instances, levels of uncertainty exist for the player at the moment of choice as to the future outcomes of the game. Whether that uncertainty arises from a lack of knowledge of another player's future actions or one's own future actions is inconsequential.

of delays and to obtain information on the amount of time participants took to make choices, that is, to increase internal validity.

Method

Participants

Fifty undergraduate students (23 males and 27 females) from the Carnegie Mellon University subject-pool participated in this experiment.

Apparatus and Procedure

All experimental stimuli were presented on an iMac computer with a 14 in. screen (13 in. viewable) using the cT programming language. Participants' choices were made using mouse clicks at the appropriate areas of the screen.

Participants were first asked a series of demographic questions including gender, age, SAT scores, etc. Following this, participants were shown an instruction screen that read as follows:

> You will be playing a game on the board shown on the left [See Figure 1] in which you will be using keys to open doors. These keys and doors will be Red and Green. Red keys open Red doors. Green keys open Green doors. At any given time you will possess a single key. This key can be used to open the appropriately colored door. Whenever you use a key to open a door, you will give up that key. Upon opening a door, you will receive the number of points contained in the box with the door, and a new key. Choices will be made by pressing buttons which will appear on the doors. While you are playing the computer will always show you the last choice you made and the number of points you have earned. You will be given a Red key to begin the game, but after that, the key you have will depend on your actions. Throughout the game, the board will look exactly as it does now. Your goal in playing this game is to earn as many points as you can.

A series of screens followed giving examples of the rules. Post-experimental questioning revealed that all participants understood the rules and procedures of the game upon reading the instructions.

Following the instructions, participants began the actual choice procedure. To avoid procedural errors during the choice procedure, only the currently valid choices (the top doors if the participant possessed a red key, the bottom two if green) were available for choice.

Participants were given a red key to begin the game and were allowed to choose between the top two doors (choose between 15 and 20 points). Regardless of the row in which the participants were currently choosing, choice for the greater points always led to the bottom of the board on the following trial and choice for the smaller points always led to the top of the board on the following trial. The "solution" to this game is such that choice for the smaller amount at the moment always led to more on future trials.

Participants made 100 choices on the screen shown in Figure 1. Participants were able to see the game board at all times, as well as the key they currently possessed, their previous choice, and the total points they had earned. An inter-trial interval (ITI) occurred between each choice (default value 3 s). During the ITI all text was removed from the screen as were the current key and previous choice information. During the ITI the key from the chosen square moved down to the current key position and the points from the chosen square were moved to the points bucket. Following the ITI of the 50th choice, a 4-min filler task (drawing) was employed. Following the filler task, participants made the final 50 choices.

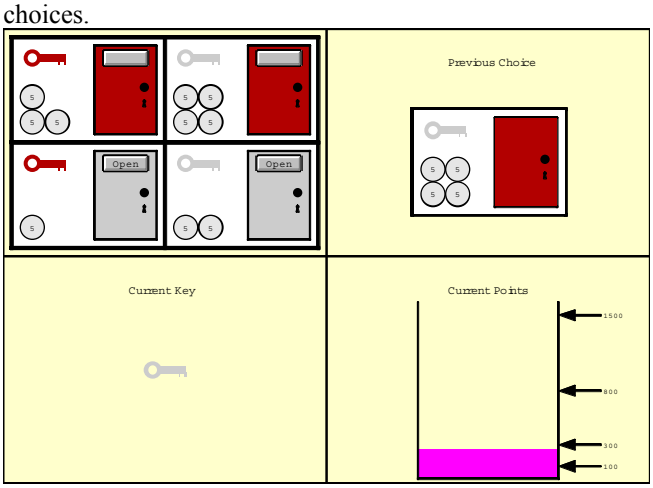


Figure 1: Screenshot of the game. Note: The top two doors are red; the bottom two doors are green; the left compartments contain red keys; the right compartments contain green keys.

Other non-prisoners' dilemma work suggests two features of the game that can impact participants' choices. These are reward amount and ITI. Within the limits of the equations presented earlier, these features can be manipulated. While the ratio of the rewards is the most important factor in choice, absolute levels of the rewards also has a considerable effect (Rachlin, Brown, & Cross, 2000). With regard to the ITI, previous work has shown that engendering commitment to a choice increases cooperation (Rachlin, 1991; Stuart, 1967). One way to accomplish this is by delaying the time until the choice takes place, in other words, increasing ITI.

Experiment 1a manipulated ITI as a between-groups factor to take the values of 3, 6, or 9 s. Additionally, within the 3 s level of ITI, the absolute level of reward was manipulated by a factor of five as a between-groups factor by writing either 5's in the circles of each choice box (as shown in Figure 1) or by writing 1's in the circles.

Results and Discussion

Performance on the iterated prisoner's dilemma game is a function of learning. As such, typical measures of

learning, including both latency and performance, were measured. Performance was measured with the percent of trials on which participants cooperated (chose the options that contained fewer points, top left or bottom left box). The first two results address whether cooperation was influenced by ITI or the absolute level of the reward.

An independent-measures ANOVA was performed across the three levels of ITI. Cooperation over all 100 trials did not significantly differ across ITI levels of 3 s (M = 86.4%, sd = .166), 6 s (M = 87.1%, sd = .170), and 9 s (M = 86.1%, sd = .162), $F(2, 47) = .01$, $p > .05$. The absolute level of reward variable was varied only within participants using an ITI of 3 s. An independent-measures t-test with these participants revealed that having 5's in the circles (M = 86.7%, sd = .157) did not engender significantly different cooperation than having 1's (M = 85.8%, sd = .191), $t(28) = .13$, $p > .05$. Because neither of the manipulated factors had a significant impact on cooperation further testing collapsed across these factors.

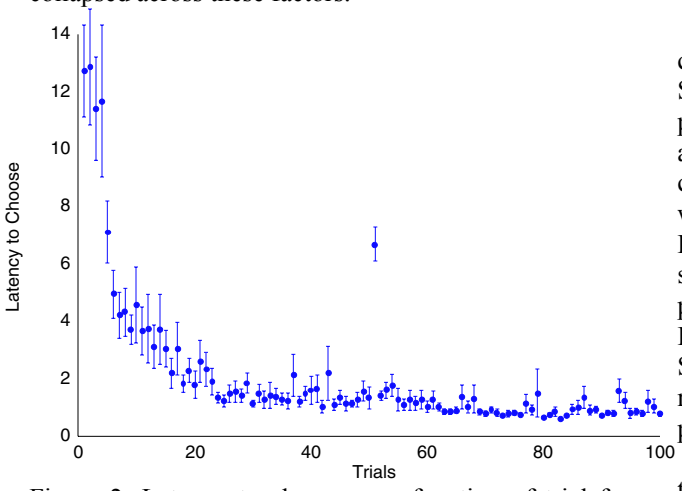


Figure 2: Latency to choose as a function of trial for Experiment 1a. Note: Bars represent ± 1 SE; the single extreme data point at trial 51 occurred immediately following the filler task.

The latencies to choose decreased in an inverse relationship with trials (see Figure 2). The best fitting curve for this data (r^2 = .823) was:

$$\text{Latency(Trials)} = 1.7026/\text{Trials}^{.6636}$$

Cooperation by the participants across the 100 trials of Experiment 1a increased logarithmically (see Figure 3). The best fitting curve for this data (r^2 = .634) was:

$$\text{Cooperation(Trials)} = 7.203*\ln(\text{Trials}) + 60.25$$

The previously mentioned manual experiment (Brown & Rachlin, 1999) was run for only 40 trials. An independent-measures t-test using only the first 40 trials of the present experiment showed that the participants in Experiment 1a (n = 50; M = 81.6%, sd = .170) cooperated significantly more than the participants in the previous experiment (n = 20; M = 53.75%, sd = .097), $t(59) = 8.61$, $p < .001$.

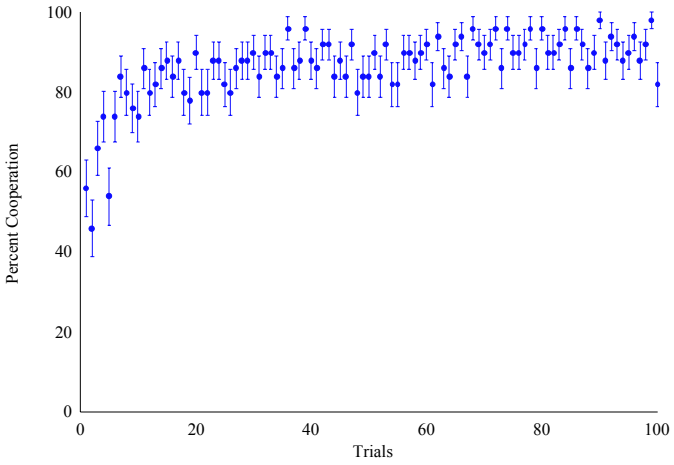


Figure 3: Percent cooperation as a function of trials during Experiment 1a. Note: Bars represent ± 1 SE.

Experiment 1a revealed a ceiling effect on cooperation that had not been seen in previous work. Several possible explanations for this effect exist. First, previous work had been performed using physical apparatus which may be viewed differently from a computer. Second, Experiment 1a was 100 trials long whereas previous work had continued for only 40 trials. However, the ceiling effect was already beginning to show by the 40th trial of Experiment 1a. Third, and perhaps most importantly, the participants in Experiment 1a were extremely analytical (average SAT-M score over 700). Post-experimental questioning revealed that participants treated the experiment as a problem to be solved mathematically.

Experiment 1b was created primarily to investigate the final concern. However, Experiment 1b was also created to increase the external validity of the game.

Experiment 1b

When making choices in life, one rarely knows for certain the consequences of those choices. In fact, it is only through prior experience that we have any idea of future consequences. This experience-based learning was missing from Experiment 1a and was created in Experiment 1b by removing the keys from the boxes.

Method

Participants

Thirty undergraduate students (19 males and 11 females) from the Carnegie Mellon University subject-pool participated in this experiment.

Apparatus and Procedure

Experiment 1b used exactly the same apparatus as Experiment 1a. The ITI in Experiment 1b was 3 s and the circles had 1's written in them. Experiment 1b differed from 1a in that some of the information which had previously been available to the participants was removed from the game. Namely, participants no longer had access to information about their actions on the previous trial and the locations of the keys in the game board (representing future consequences). Additionally, ten of the participants provided a verbal protocol while performing the experiment.

Results and Discussion

An independent-measures *t*-test on the overall percent cooperation showed that participants providing a verbal protocol (M = 83.2%, sd = .172) did not significantly differ from those not providing it (M = 79.9%, sd = .223), $t(28) = -0.41$, $p > .05$. Because of this, further analyses collapsed across this factor.

However, protocol data revealed that participants were still "solving" the game as a math problem. Participants were adding together the outcomes of different combinations of multiple trials, comparing these totals, then merely selecting the combination with the highest total. These participants were again highly analytical (SAT-M average over 700).

In much the same way as Experiment 1a, latency decreased across trials in an inverse relationship. The best fitting curve for this data (r^2 = .739) was:

$$\text{Latency(Trials)} = 8.77/\text{Trials}^{.5468}$$

The cooperation in Experiment 1b increased across trials logarithmically as Experiment 1a. The best fitting curve for this data (r^2 = .653) was:

$$\text{Cooperation(Trials)} = 15.18*\ln(\text{Trials}) + 25.75$$

A 2 X 10 (Experiment X Ten Trial Blocks) mixed-factorial ANOVA was performed. The overall cooperation in Experiment 1a (M = 86.5%, sd = .163) was not significantly different from Experiment 1b (M = 81%, sd = .205), $F(1, 78) = 1.73$, $p > .05$. The cooperation across the ten ten-trial blocks (M's = 59.9%, 81.3%, 82.4%, 86.4%, 85.9%, 82.9%, 89.4%, 92%, 92.0%, 92.1%; sd's = .263, .242, .240, .229, .240, .216, .204, .185, .174, .166) was significantly different, $F(9, 702) = 53.59$, $p < .001$. A trend-analysis revealed a significant linear component, $F(1, 79) = 76.64$, $p < .001$, suggesting that steady learning occurred across the experiments. Additionally, the interaction of Experiment and trial block was significant, $F(9, 702) = 7.66$, $p < .001$. Planned comparisons showed that the essential differences in Experiments 1a and 1b occurred during the first [$F(1, 78) = 15.86$, $p < .001$] and sixth [$F(1, 78) = 8.19$, $p < .01$] blocks of ten trials in which the participants' in Experiment 1b cooperated less.

Procedurally, Experiment 1a and 1b differed in a quantitative manner. The task was made slightly more difficult by removal of the keys (information about future consequences) and the previous trial reminder (information about past behavior). This quantitative change in task difficulty created only a slight difference in overall cooperation, with its effects felt mainly in the first and sixth block of trials (immediately following the filler task).

Throughout Experiments 1a and 1b, all participants (N = 80) "solved" the game in a sudden fashion using one of several strategies. The first strategy was total cooperation, which was achieved by selecting the top-left door continually. The second strategy, which yielded approximately 66% cooperation, involved choosing the top left door, then the top right door, then the bottom left door, then repeating the sequence. The third common strategy, which created approximately 50% cooperation, involved choosing the top right door, then the bottom left, over and over. Participants using the first strategy (n = 61) tended to do so during the early trials of the experiment (M = 12.3, sd = 15.1). Participants using the second strategy (n = 11) tended to begin it late (M = 40.1, sd = 24.7). Participants implementing the third strategy (n = 8) did so in between (M = 29, sd = 26.9). Once participants moved into one of these strategies, they stuck with them nearly exclusively. Strategy onset was defined using the following two rules. First, the behavior for at least 10 trials following the point of strategy onset must be consistent. Second, the participant must have followed this strategy throughout the game.

A one-way repeated measures ANOVA was performed on the latency to choose on the five trials surrounding the strategy onset (two trials prior to onset, actual onset, and two trials following onset). Eleven participants were removed from this analysis (all using the pure cooperation strategy) because their strategy onset occurred on the 1st or 2nd trial such that they provided no data for trials prior to onset (if included, the effect is more pronounced). These latencies as a function of position, which can be seen in Figure 4, were significantly different, $F(4, 272) = 3.58$, $p < .01$. A planned comparison revealed that the latency on the trial immediately following strategy onset (M = 8.45 s, sd = 15.9) was significantly longer than all other trials averaged together (M = 4.28 s, sd = 7.5), $F(1, 68) = 6.39$, $p < .05$.

Prior to the onset of a steady strategy, verbal protocol data revealed that participants' comments tended to focus on the outcome of the immediate trial. The trial on which strategy onset occurred was treated no differently from previous trials. However, on the trial following strategy onset, the participants would stop and reconsider the entire game, focusing on the game as a "whole", hence, the increased latency on the trial

immediately following strategy onset. The results are particularly striking given that earlier trials tend to have a longer latency (see Figure 2 and the downward slope among 4 of the 5 points in Figure 4). This aspect of the data would have the effect of negating the increased latency on the trial following strategy onset.

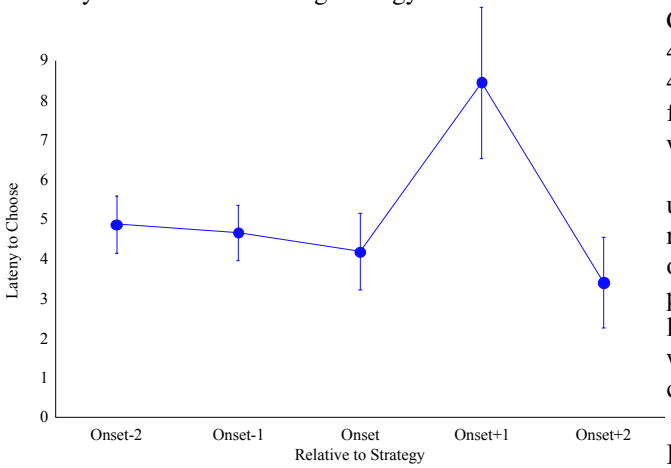


Figure 4: The latency to choose on the trials surrounding strategy onset in Experiment 1a and 1b. Note: Bars represent ± 1 SE.

Experiment 2

Due to the ceiling effects experienced in both Experiments 1a and 1b, participants' treatment of the game as a mathematical problem to be solved was addressed. Real-life choices rarely can be solved mathematically. It is often impossible to verbalize why option A was chosen over option B, it may simply "feel" better. This intuitive nature of decision-making was sought by creating two probability combinations such that reinforcement was unknown in advance rather than deterministic as it had been in Experiment 1.

Method

Participants

Forty undergraduate students (27 males and 13 females) from the Carnegie Mellon University subject-pool participated in this experiment.

Apparatus and Procedure

Experiment 2 used the same apparatus as Experiments 1a and 1b. The ITI was 3 s and the circles had 1's written in them. It will be remembered that in Experiment 1 (see Figure 1) the game board contained 3, 4, 2, and 1 circles in the different compartments. In Experiment 2 all compartments contained 5 circles. However, in Experiment 2, upon making a choice, the receiving of points was probabilistic. If a participant chose any given door, they may or may not receive the 5 points. The probabilities used were manipulated. Participants in Group 1 (n = 20) received the 5 points 60%, 80%, 40%, or 20% of the time (starting in the top left and going clockwise). This created expected values for these choices that were 3, 4, 2, and 1 (i.e. 5 points received 60% of the time has an expected value of 3), exactly as they were in Experiment 1. Participants in Group 2 (n = 20) received the 5 points 80%, 100%, 40%, and 20% of the time creating expected values of 4, 5, 2, and 1. Due to an apparent lack of asymptote following 100 trials, 12 of the participants in Group 2 were given 100 additional trials for a total of 200.

The previously listed probabilities were generated using a true probability mechanism, However, several rules were employed which ensured both that the obtained probabilities approximated the programmed probabilities at both the local and global level and that long strings of wins or losses (such that the string would be expected to occur less than 5% of the time by chance) were avoided.

Results and Discussion

A 2 X 10 (Group X Block of ten trials) mixed-factorial ANOVA was performed on the cooperation during the first 100 trials. Group 1 (M = 53.5%, sd = .143) did not cooperate significantly less overall than Group 2 (M = 58%, sd = .106), $F(1, 38) = 1.28$, $p > .05$. The cooperation across the ten ten-trial blocks did not significantly differ, $F(9, 342) = .36$, $p > .05$. The interaction of Group and block was not significant, $F(9, 342) = 1.2$, $p > .05$.

In a similar manner seen in the differences between Experiments 1a and 1b, the slight difference in results for Groups 1 and 2 seemed to reflect a quantitative change in task difficulty. Group 2's more divergent probabilities, and hence expected values, created more cooperation.

A single-sample t-test on the overall cooperation for Group 1 showed that it did not significantly differ from random responding (50%), $t(19) = 1.08$, $p > .05$. A second single-sample t-test on cooperation during the final ten-trials (when responding is most stable, M = 55.5%, sd = .233) revealed the same lack of difference, $t(19) = 1.06$, $p > .05$.

Perhaps the probabilities of reinforcement used with Group 1 were too subtle even though the expected values were identical to the points used in Experiment 1. Inspection of the data shows virtually no change in cooperation from the beginning to the end of the experiment. The effects of continuing the experiment beyond 100 trials remain to be seen.

A single-sample t-test on the cooperation for Group 2 during the first 100 trials (n = 20) showed that it was significantly higher than chance, $t(19) = 3.35$, $p < .01$. However, cooperation during the 10th block of ten trials

(M = 58.5%, sd = .223) was not significantly higher than expected by chance, $t(19) = 1.7$, $p > .05$.

A single sample t-test on cooperation over all 200 trials (M = 63.4%) for those that received them (n = 12) showed that it was significantly higher than chance, $t(11) = 3.33$, $p < .01$. Likewise, cooperation during the final ten-trial block (M = 73.3%, sd = .227) was significantly higher than chance, $t(11) = 3.56$, $p < .01$.

The cooperation of Group 2 continued a slow growth throughout the first 100 trials of the experiment. This growth also continued through the second 100 trials. One wonders where this group's cooperation would asymptote.

The latency data of Experiment 2 mirrored the inverse relationship seen in both Experiments 1a and 1b. Latency decreased across trials in an inverse relationship. The best fitting curve for this data, which created an r^2 of .896, was:

$$\text{Latency(Trials)} = 5.52 / \text{Trials}^{.4186}$$

An examination of the relationship of cooperation across trials for the participants in Group 1 showed that no function could account for more than 2% of the variance. The relationship between cooperation and trials was slightly better for Group 2. A logarithmic function captured the greatest amount of variance in this relationship creating an r^2 of .151:

$$\text{Cooperation(Trials)} = 5.11 * \ln(\text{Trials}) + 41.52$$

A similar analysis of strategy, as was done in Experiment 1, was attempted for Experiment 2. Even when using a lenient definition of strategy onset, only 16 of the 40 participants were classified as having employed a steady strategy. A one-way repeated measures ANOVA on the latency to choose on the trials surrounding strategy onset revealed no differences, $F(4, 60) = .89$, $p > .05$.

General Discussion

When the internal validity of the game was increased (prior experimentation to the present experiments), participants' ability to cooperate with themselves increased substantially (though population differences may have had a large impact on this change). As the complexity, external validity, and face validity of the game increased within these experiments (both quantitatively and qualitatively), participants' ability to cooperate with themselves decreased.

Participants' behavior in Experiments 1a and 1b was characterized by a sudden change: at one moment focusing on the current trial, at the next moment focusing on the whole game. This "insight learning" showed itself in the strategy results. There was some moment when each participant stopped responding semi-randomly and began responding according to a strategy. In Experiment 2 behavioral change occurred gradually, though slightly faster for Group 2. This "trial and error" learning was shown in two ways. First, by the absence of strategy results. Second, by the participants' use of a win-stay/lose-shift approach.

The procedural differences between Experiments 1a and 1b created a quantitative change in behavior, 1b was slightly more difficult than 1a. Participants in both experiments "solved" the problem in a moment of insight. The procedural differences between Group 1 and Group 2 in Experiment 2 created quantitative differences in behavior, with the more divergent expected values in Group 2 creating more cooperation. Both groups appeared to use a form of trial and error learning. However, the change from Experiment 1 to 2 was qualitative. Participants solved these experiments using completely different approaches.

The insight shown in Experiment 1 occurred only after the participants had started using a strategy. On the trial following strategy implementation, the participants stopped and viewed the entire experiment, adding together the points from various combinations of moves. Because no verbal protocols were used in Experiment 2, it is impossible to know for sure how these participants viewed the problem. Perhaps the participants viewed the game as they would a real life problem. Picking one option "feels" better than picking another. Perhaps a simpler explanation based on probability matching may be a better explanation.

References

Axelrod, R. (1987) The Evolution of Strategies in the Iterated Prisoner's Dilemma. In Lawrence Davis (ed.), *Genetic Algorithms and Simulated Annealing*. London: Pitman.

Brown, J. C. & Rachlin, H. (1999). Self-control and social cooperation. *Behavioural Processes*, *47*, 65-72.

Rachlin, H. (1991). *Introduction to Modern Behaviorism* (3rd ed.). New York: W. H. Freeman and Co.

Rachlin, H., Brown, J., & Baker, F. (2001). Reinforcement and punishment in the prisoner's dilemma game. In D. L. Medin (ed.), *The Psychology of Learning and Motivation-Advances in Research and Theory: Vol. 40*. New York: Academic Press.

Rachlin, H., Brown, J., & Cross, D. (2000). Discounting in judgments of delay and probability. *Journal of Behavioral Decision Making*, *13*, 145-159.

Rapoport, A. & Chammah, A. M. (1965). *Prisoner's Dilemma-A Study in Conflict and Cooperation*. Ann Arbor, MI: The University of Michigan Press.

Stuart, R. B. (1967). Behavioral control over eating. *Behavior Research and Therapy*, *5*, 357-365.

Mice Trap: A New Explanation for Irregular Plurals in Noun-Noun Compounds

Carolyn J. Buck-Gengler (Carolyn.Buck-Gengler@Colorado.edu)
Department of Psychology and Institute of Cognitive Science
University of Colorado at Boulder
Boulder, Colorado 80309-0345 USA

Lise Menn (Lise.Menn@Colorado.edu)
Department of Linguistics and Institute of Cognitive Science
University of Colorado at Boulder
Boulder, Colorado 80309-0295 USA

Alice F. Healy (Alice.Healy@Colorado.edu)
Department of Psychology and Institute of Cognitive Science
University of Colorado at Boulder
Boulder, Colorado 80309-0345 USA

Abstract

In an experiment eliciting noun-noun compounds, participants were more likely to produce plural nouns in the first position (e.g., *mice trap*) when presented with an irregular plural in the stimulus (e.g., *a trap for catching mice is a* _____) than when presented with stimuli containing regular plurals (e.g., *a trap for catching rats is a* _____). When they did produce a normatively correct singular (e.g., *mouse trap*) in response to a stimulus with an irregular plural, response time was longer than it was for producing a singular response to stimuli containing singulars or regular plurals. This finding suggests a priming-based processing problem in producing the singulars of irregular plurals in this paradigm. Such a problem is likely also to be present in young children, which would explain their production of forms like *mice trap* (Gordon, 1985; Pinker, 1994) without requiring recourse to the hypothesis that they have innate grammatical knowledge.

Introduction

It has been observed that irregular nouns and verbs often behave quite differently than their regular counterparts. A case in point is noun-noun compound formation in English. Pinker (e.g., 1994, pp. 146-147) and others have pointed out that in general, the first noun of a noun-noun compound must be singular, but that plurals of irregular nouns seem to be more acceptable in that position. For instance, one would not say **rats catcher* or **toys box*, even when talking about more than one rat or toy. However, *mice catcher* is far more acceptable. The typical explanation given is that this follows from the theory of Level Ordering of morphology (Kaisse & Shaw, 1985; Kiparsky, 1982, 1983).

In this theory, production of compounds (or at least, novel compounds) proceeds at several "levels." At Level 1, a base form (for almost all English nouns, the singular) or another memorized form (such as the irregular plural) is retrieved from the mental lexicon; at Level 2, compounds are formed; at Level 3, after compound formation, regular affixes such as the regular plural are added. If this production schema is taken to represent a sequence of real-time processes or a fixed mode operation of a mental grammar, the normative English pattern dispreferring **rats trap* is explained by saying that the regular plural *rats* is created too late (at Level 3) to be placed inside a compound (at Level 2). However, irregular plurals, being retrieved from memory at Level 1, are easily incorporated during compound formation. This theory, in its general form, successfully schematized a wide variety of data in English and other languages, although by now it has been shown to have serious limitations (Bauer, 1990; Fabb, 1988).

Gordon (1985) explored compound formation with children aged 3-5, and induced them to produce compounds containing plurals when presented with irregular nouns, but only singulars when primed with regular nouns. He took this result as support for the idea that level ordering in grammar must be innate, because, as he demonstrated, children are rarely exposed to irregular plurals inside compounds, so they cannot induce the rule permitting irregular plurals inside compounds from their input.

However, there is another way to formulate the English pattern: English compounds obey the (soft) constraint that the first element is singular, regardless of whether the semantics has stimulated the retrieval of a plural referent (e.g., *toys*, *cookies*). Thus, if a compound such as *cookie jar* is called for, the retrieved *cookies* must be made singular before proceeding. Young children have plenty of examples of this pattern in their early years (e.g., *toy box*, *raisin box*, *jelly bean jar*) and could well have adduced such a rule by age 3.

Using this second way of constructing compounds, we suggest a processing difficulty explanation for the experimentally observed behavior: It is harder to produce a singular when primed by an irregular plural (e.g., to produce *mouse* after just having heard or read *mice*) than it is to

produce a singular when primed by a regular plural (e.g., to produce *rat* just having heard or read *rats*), for reasons that will be discussed later. This difficulty should show up in two ways: first, the already observed predilection of both adults and children to produce the irregular plural noun in compounds more often than the regular plural (predicted by both explanations), and second, a longer time to produce such a compound when given an irregular plural as input (predicted by our explanation).

To examine these predictions, we constructed an experiment to elicit noun-noun compounds and obtain response times. The experiment was similar in some ways to that of Gordon (1985), but differed in several important features. As in Gordon's experiment, the compound was elicited by prompting with some form of the words that were intended to be used in the compound. The differences were as follows: First, adult speakers rather than children were used. Second, response times were recorded by a computer, and therefore the materials were also presented by the computer. Third, we included the set of regular and irregular nouns that Gordon used, but augmented it in three ways: (a) Both the singular and the plural forms of each noun were used as stimuli; (b) three additional sets of regular and irregular words were added to provide for more stimuli; and (c) in addition to semantically matched regular nouns, a set of regular nouns matched for form (length, frequency, and initial phoneme class) was included. Participants were trained on the individual words, then responded to fill-in-the-blank sentences as quickly as possible with compounds that included one of the target words as the first word and a container as the second word.

Method

Participants

16 University of Colorado psychology students participated in the experiment. All were native speakers of English. Participants were assigned to counterbalancing groups according to a fixed rotation on the basis of their time of arrival at the experiment.

Apparatus

Stimuli were presented on an Apple iMac computer using the PsyScope software (Cohen, MacWhinney, Flatt, & Provost, 1993) and a button box/voice key. Responses were recorded on a cassette tape recorder via a head-mounted microphone through a y-jack.

Procedure

There were two parts to the experiment, training and test. The entire experiment took approximately 35-45 minutes.

Training The training part of the experiment familiarized participants with the task and the set of words used in the experiment. During training, participants were shown words one at a time. Each word that was to be used in the test portion of the experiment was shown twice, in two different pseudorandom orders, for a total of 152 trials. Participants were instructed to name the word as quickly and clearly as possible.

For each trial, first a black dot was displayed. When participants were ready for the trial, they pressed the space bar, causing the dot to disappear and the word to be displayed. When the participant responded, the computer registered the response via the voice key; the time was recorded; the word was erased; a red asterisk was briefly displayed as feedback that recording had taken place; and then the black dot signaling the next trial was displayed. Response time was measured from the time the word was displayed to the time the response triggered the voice key. Self-pacing of trials gave participants control over the pacing of the experiment, and also allowed a participant to postpone a trial momentarily if he or she felt a need to cough or make other noise that could disrupt the voice key. The feedback asterisks also trained participants to speak loud enough to trigger the voice key.

Test The second part of the experiment was similar to the first, except that in addition to single-word trials, there were also complex trials involving reading a sentence of the form "a JAR containing COOKIES is a _____" and filling in the blank. This part began with examples to show the participants what the complex fill-in-the-blank stimulus sentences would be like, and instructing them to fill in the blank based on the sentence they saw. The first example showed the participant what a "typical" answer would be; the participant then practiced on seven more examples by answering what they thought the answer should be and then repeating what the computer printed in red as the "typical" answer. (The example sentences are described in the Materials section.)

In this part of the experiment, the self-pacing was done as in the training phase. There were 4 blocks in this part of the experiment. Each block had 60 complex trials mixed with 90 single-word trials; together with 10 practice trials (different from the 8 example trials) before the first block, there was a total of 610 trials in the second part of the experiment. The order of the four blocks was counterbalanced using a balanced Latin Square design, resulting in four different groups of participants.

Materials

Target Nouns There were three types of target nouns: irregular nouns, semantically matched regular nouns ("semantic match"), and form matched regular nouns ("form match"). Five of the irregular nouns and their semantic matches were taken from Gordon (1985). These lists were augmented with three more nouns for greater generalizability across items. To draw attention away from the irregular nouns and their semantic matches, a non-semantically related noun was matched with each irregular noun. This set of nouns was chosen such that each matched regular noun was similar in length and frequency to the irregular noun, and also started with a phoneme that had acoustic onset characteristics similar to the first phoneme of the irregular noun. In addition, six more nouns and their plurals (two irregular nouns and four regular nouns) were

Table 1: Target nouns by type.

	Regular noun	
Irregular noun	Semantic match	Form match
mouse*	rat*	nail
tooth*	bead*	tape
foot*	hand*	cent
goose*	duck*	bell
man*	baby*	letter
louse	fly	knight
child	doll	chain
ox	horse	ax

*From Gordon (1985)

used for filler trials. The complete set of target nouns (singular form) is presented in Table 1. Targets in the experiment included both singular and plural forms.

The three new irregular nouns were chosen to be imageable and concrete. Semantic match nouns were chosen with criteria similar to those of Gordon (1985). Form match nouns, in addition to fitting frequency and length constraints, also fit the acoustic onset match criterion. Frequency and word length were equated across irregular, semantic match, and form match lists.

Individual Stimuli Individual stimuli were either single-word or complex. Single-word stimuli consisted of a single word chosen from the target nouns, containers, fillers, and words from the practice trials of the second part of the experiment. Complex stimuli consisted of fill-in-the-blank sentences of the following format: "a CONTAINER filled with (a/an) TARGET NOUN is a/an _____". Table 2 lists the containers and the verb associated with each container; TARGET NOUN was replaced with the target noun for that trial. The goal was to elicit a noun-noun compound using some form of the target as the first noun and the container as the second noun. The container and target noun were in upper case and the rest of the sentence was in lower case.

Each of the 48 target nouns and 12 filler nouns was combined with each of the 4 containers for a total of 192 target and 48 filler complex stimuli. Each target and filler noun occurred once in each block; the containers were distributed over the 4 blocks using a modified balanced Latin Square pattern.

In addition, there were eight example stimuli (all complex) and 10 practice stimuli (six single-word and four complex) at the beginning of Part 2 of the experiment, which were included (a) to ensure that the participants were not merely reversing the words, and (b) to give participants experience with the second part of the experiment before beginning the experimental trials. The eight example trials alternated between using mass nouns as the target nouns (because mass nouns do not have a plural/singular distinction, and thus would not have a distinct plural form; i.e., *rice jar*, *dough pan, fish glass, soup pot*) and extremely common noun-noun collocations. These examples were intended to induce the participants to realize that changing the form of the target noun was allowable (i.e*., bird cage, tool chest, coin purse, egg carton*). In these example trials

Table 2: Containers and associated verbs.

CONTAINER	Verb (*replaces* "filled with"*)*
bowl	containing
box	for transporting
crate	for carrying
tub	holding

the plural form of the target noun was presented, but the expected or typical answer was the singular form. Common noun-noun compounds were also chosen for the four complex practice stimuli (i.e., *flower vase, pencil case, chicken coop, cookie jar*). As with the example stimuli, the plural form of the target noun was presented in the complex practice stimuli. The target nouns and containers from these complex examples were included as single-word trials in Part 1. The single-word trials in the practice stimulus set were taken from the eight example stimuli used at the beginning of the test part of the experiment.

Stimulus lists As noted earlier, in Part 1 all 48 experimental target words, 12 filler words, 4 container words, 4 compound practice containers, and 8 compound practice target words (both singular and plural forms) were presented twice, in two different pseudorandom orders. All participants saw the same order of stimuli.

In Part 2 (test), single-word trials and complex trials were intermixed. First, the complex trials for each block were pseudorandomly ordered. Within each subblock of 12 complex trials there were two words of each combined type (noun type x grammatical number). Preceding each complex trial was either 0 or 1 container single-word trial, and either 0 or 2 target word single-word trials. There were no more than 2 complex trials (i.e., with 0 intervening single-word trials) in a row.

Design

The dependent variables measured were the response time (RT) for each compound response in ms and the proportion of complex trials with a singular first noun response (out of all usable trials as defined in the section on scoring). Each measure was analyzed with a mixed 4 x 3 x 2 analysis of variance. The first factor (between-subjects) was counterbalancing group. The second factor (within-subjects) was the noun type (irregular, semantic-match regular, form-match regular) of the target noun. The third factor (within-subjects) was the grammatical number (singular, plural) of the target noun.

A number of participants gave no plural responses to one or both of semantic match plural and form match plural stimuli. To reduce the number of participants whose data therefore would have to be eliminated from the RT analyses, the third analysis was conducted over the combined results of the regular nouns. Collapsing these categories was further justified by post-hoc tests that determined that there was no significant difference between the results of the two sets of regular nouns. In this analysis, the dependent variable was RT, which was analyzed with a mixed 4 x 2 x 2 analysis of variance. The first factor (between-subjects) was counterbalancing group. The second factor (within-subjects)

was the noun type (irregular, regular) of the target noun. The third factor, although not properly an independent variable, was response type (singular, plural) and was also within subjects.

Scoring

Two independent scorers listened to each tape. Each trial, single-word or complex, was scored as singular, plural, or other. Problems that would affect the RT as recorded by the voice key, such as repeating a word more than once, coughs or other noises, giving the complex answers in the wrong order, and so forth, were noted. Discrepancies between the two scorers were resolved by a third listener. RTs for problem trials and trials with responses categorized as "other" were discarded, and those trials were not used in RT analyses. When a participant responded to a complex stimulus first with one complete answer and then with another, it was scored with the first response.

Results

All results are based only on complex trials that were not excluded due to problems, as discussed above.

Proportion of Trials with Singular First Noun Responses The first analysis examined how many responses used a singular first noun, no matter what the grammatical number of the stimulus. As expected, there was no effect of counterbalancing group, nor did this factor interact significantly with any other factor. Whereas approximately 93.8% of all semantic match trials and 94.6% of all form match trials were responded to with the singular form of that noun, only 64.1% of all trials with an irregular target noun were responded to with the singular form (thus, 35.9% were responded to with the plural form). This difference was significant, $F(2,24) = 75.31$, $MSE = .013$, $p < .001$. Of trials with a singular target noun, 99.1% of the responses were also singular, whereas only 69.2% of the trials with a plural target noun were responded to with the singular form. This difference was also significant, $F(1,12) = 46.94$, $MSE = .046$, $p < .001$.

What is interesting is the interaction between these two factors. As can be seen in Figure 1, when the trial contained an irregular plural target noun, the response was the singular form only 30.0% of the time. For all five of the other combinations, the singular form was used between 88.8% of the time (for the regular plural target nouns, including both semantic and form match) and 99.1% of the time (for the singular nouns of any type). This interaction was significant, $F(2,24) = 70.534$, $MSE = .012$, $p < .001$. This result confirms earlier findings, notably those of Gordon (1985), that irregular plurals are readily produced as the first noun of noun-noun compounds in response to this type of elicitation frame.

RTs to Singular First Noun Responses This analysis looks at the time to respond with singular *mouse box* when shown either "a BOX for transporting a MOUSE is a _____" or "a BOX for transporting MICE is a _____", compared to the time to respond when regular forms were

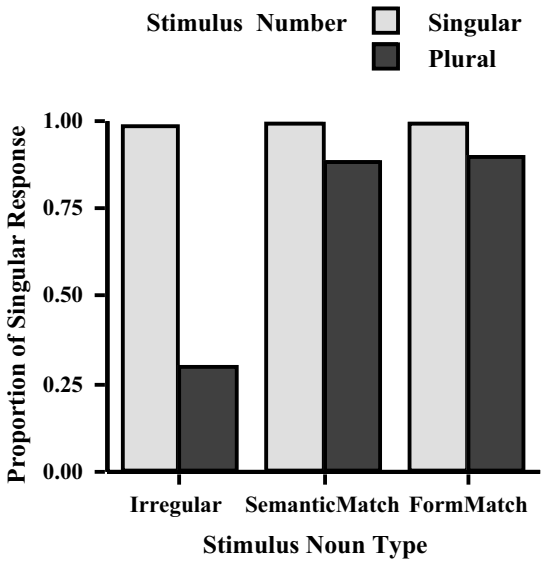


Figure 1: Proportion of singular response by noun type and grammatical number of the stimulus target word.

shown. This analysis includes only those trials in which the response was singular; thus fewer trials contributed to the irregular plural cell than to the other cells (as seen in Figure 1). Four participants responded with a plural 100% of the time when given an irregular plural stimulus. Thus, they had zero singular responses, and no mean RT could be computed for that cell. As a result, the data from those four participants were excluded from this analysis.

As expected, there was no effect of counterbalancing group, nor did this factor interact significantly with any other factor. RTs for the semantic match trials (796 ms) and the form match trials (799 ms) were significantly faster than for the irregular trials (894 ms), $F(2,16) = 6.18$, $MSE = 13179$, $p = .010$. RTs for singular-stimulus trials (787 ms) were significantly faster than for plural-stimulus trials (872 ms), $F(1,8) = 13.61$, $MSE = 10430$, $p = .006$. Most interestingly, the interaction between these factors, seen in Figure 2, was also significant, $F(2,16) = 6.90$, $MSE = 8075$, $p = .007$. As can be seen in the figure, it is the irregular plural stimuli that have the slowest RT; all other forms were responded to much more quickly.

An additional finding is that there was no significant difference between the semantic match and form match regular nouns, as can be seen in both Figure 1 and Figure 2, and confirmed by post-hoc tests. This result means the differences between regular and irregular nouns do not depend either on semantic or form similarity.

RTs to Trials with Plural Stimuli This analysis looks at the time to say *mouse box* or *mice box* when shown "a BOX for transporting MICE is a _____", compared to the time to say *rat box* or *rats box* when shown "a BOX for transporting RATS is a _____". In this analysis, both singular and plural responses were examined for all trials with a plural stimulus noun. As noted previously (and as can be inferred from Figure 1), the rate of plural response to

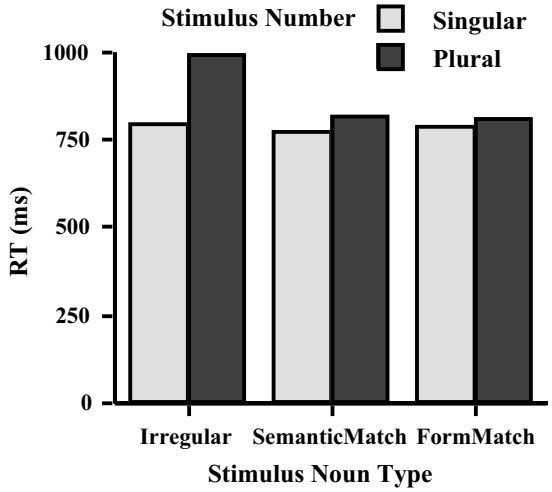


Figure 2: Response times in ms for singular responses.

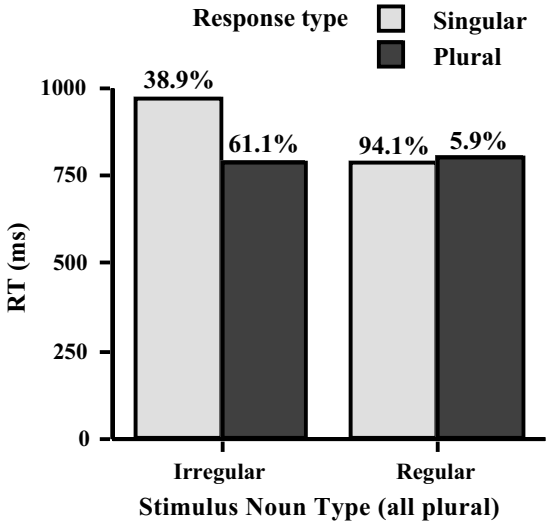


Figure 3: Response times in ms for both singular and plural responses to plural target noun stimuli. Numbers over the bars indicate the proportion of trials contributing to that cell.

regular plural stimuli was extremely low; so to increase the chance of a plural response, the two types of regular nouns were combined for this analysis. In addition to the four participants with no singular responses to irregular plural stimuli, one participant never responded to regular plural stimuli with a plural, and no mean RT could be computed for that cell. Thus, the data from these five participants were excluded from this analysis.

As expected, there was no effect of counterbalancing group, nor did this factor interact significantly with any other factor. In this analysis, neither the main effect of noun type nor the main effect of response type were significant. However, the interaction between these factors was significant, $F(1,7) = 6.36$, $MSE = 14120$, $p = .040$. As can be seen in Figure 3, there was no RT difference in producing a singular or a plural response to a regular plural stimulus or a plural response to an irregular plural stimulus. This can be thought of as the baseline time for responding. However, when participants were presented with an irregular plural stimulus to which the singular response was eventually made, the time was much longer.

Discussion

The proportion of singular response in this experiment showed a pattern similar to that found by Gordon (1985); that is, that when the participant is given an irregular plural noun in the stimulus, the response could be either singular or plural, but with all the other combinations of number and noun type, the response was almost invariably singular. Thus, this experiment, with added controls and stimuli, serves to verify that a typical college undergraduate population gives similar results to the children tested by Gordon. (Some of our participants, like some of Gordon's children, also provided self-corrections after their plural responses: "men bowl, oops, man bowl.")

What is new here are the response time findings. Specifically, when the response actually produced was singular, it took longer to produce when the stimulus was an irregular plural noun than when it was any other combination of number and noun type. Moreover, although it does not take longer to produce *rat* than *rats* when *rats* is seen, it DOES take longer to produce *mouse* than *mice* when *mice* is seen. We suggest that the times for the plural responses in Figure 3 show the times to inhibit the just-primed plural and produce the singular form instead. A model congruent with our findings would have the following properties: Given a form like *rats*, the plural affix *-s* is automatically segmented by the hearer/reader, allowing the regular singular *rat* to be strongly primed. On the other hand, given a form like *mice*, the singular is aroused only by re-accessing the lexicon where a *mouse-mice* link is stored, which presumably is a more time-consuming process.

To be more explicit, we think something like the following is going on. When the task is to respond to "a BOWL containing MICE is a _____", *mice* is strongly activated; *mouse* is likely activated as well but somewhat later and not as strongly, allowing both *mice bowl* and *mouse bowl* to be formed and compete with/inhibit each other. The constraint preferring singulars within compounds would also add inhibition to *mice bowl*, but because *mice bowl* starts out with higher activation, it wins more often. However, sometimes *mouse bowl* does win, when it and the constraint succeed in inhibiting *mice bowl*. In the regular case, when the task is to respond to "a BOWL containing RATS is a _____", *rats* is aroused, but *rat* is aroused as well. The constraint preferring singulars in compounds inhibits *rats bowl* even more strongly than it did *mice bowl*; that inhibition, along with the competition from *rat bowl*, serve to eliminate *rats bowl* in almost all cases, resulting in an output of *rat bowl* most of the time.

A number of theories of morphological structure, including Level Ordering, would be compatible with such a processing model. Our proposal does, however, postulate that speakers recognize and segment grammatical morphemes when the language structure supports it.

To be sure, our elicitation task, designed to parallel Gordon's (1985) task, and also to permit collecting reaction time data, is distinctly non-natural. The data we have collected do not speak directly to the problem of creating a real-time model of how plurals and compounds are created in natural speech. However, in the real world the situation is also not as neatly divided into regular and irregular nouns with different behavior. In the real world we find violations of the no-plurals-inside-compounds constraint (e.g., *civil rights legislation, fireworks display, parks commissioner*). Children also hear at least four nouns of English that have no singular form (e.g., *clothes*, *pants*, *scissors*, *glasses*). They hear these nouns both alone and in compounds (e.g., *clothes basket*, *pants leg*); such compounds are well-attested in input to children in CHILDES (MacWhinney, 2000). Eventually children will discover that these nouns are special in not having corresponding singulars, but initially the nouns may be apprehended as evidence that at least some plurals may be allowed within compounds.

The cornerstone of Gordon's (1985) and Pinker's (1994) argument for innateness of level ordering in grammar was that children who had no exposure to irregular plurals inside compounds nevertheless permitted them, as did adults. What we have shown is that adult production of plurals inside compounds in this type of elicitation task is probably a consequence of the difficulty in overcoming the strongly primed irregular plural form. Presumably children in Gordon's task faced the same problem. The fact that they behave like adults need not be due to their having an adult-like rule permitting irregular plurals in compounds, but rather to their having a similar human system for processing language stimuli.

To conclude, our finding points to a processing difficulty explanation for violations of the constraint against plurals within compounds in this elicitation situation: It is more difficult and time-consuming to produce the singular form of an irregular noun when primed with a stimulus of the plural form than it is to segment the regular plural marker, at least when it is an easily removable affix like the *–s* in English, and thereby retrieve the singular form. It is likely that the same explanation also works for the children examined by Gordon (1985). Given these results, we argue that Gordon's findings do not provide support for the notion of innate grammar.

Acknowledgements

We are grateful to Erica Wohldmann and James Kole for their assistance in scoring the tape recordings, to William Bright for editorial suggestions, to George Figgs for his CHILDES searches, and to Dan Jurafsky, Todd Haskell, and K. P. Mohanan for their comments and suggestions at various stages of this project. Lise Menn presented preliminary explorations of this topic on several occasions; we are grateful to audiences at the 11th World Congress of Applied Linguistics, the Max Planck Institute for Psycholinguistics at Nijmegen, the Max Planck Institute for Evolutionary Anthropology at Leipzig, Swarthmore College, and the UCLA Linguistics Department for their helpful comments. This research was supported in part by Army Research Institute Contract DASW01-99-K-0022 and Army Research Office Grant DAAG55-98-1-0214 to the University of Colorado (Alice Healy, Principal Investigator).

References

Bauer, L. (1990). Level disorder: The case of *-er* and *-or*. *Transactions of the Philological Society, 88*, 97-110.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: An interactive graphics system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments & Computers, 25*(2), 257-271.

Fabb, N. (1988). English suffixation is constrained only by selectional restrictions. *Natural Language and Linguistic Theory, 6*, 527-539.

Gordon, P. (1985). Level-ordering in lexical development. *Cognition, 21*(2), 73-93.

Kaisse, E. M., & Shaw, P. A. (1985). On the Theory of Lexical Phonology. *Phonology Yearbook, 2*, 1-30.

Kiparsky, P. (1982). From cyclic phonology to lexical phonology. In H. v. d. Hulst & N. Smith (Eds.), *The Structure of Phonological Representations* (pp. 131-175). Dordrecht, The Netherlands: Foris Publications.

Kiparsky, P. (1983). Word-formation and the lexicon. In F. Ingeman (Ed.), *Proceedings of the 1982 Mid-America Linguistics Conference* (pp. 3-29). Lawrence, KS: University of Kansas.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk. Third Edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Pinker, S. (1994). *The Language Instinct*. New York: W. Morrow and Co.

Simulating the Evolution of Modular Neural Systems

John A. Bullinaria (j.bullinaria@physics.org)
School of Computer Science, The University of Birmingham
Edgbaston, Birmingham, B15 2TT, UK

Abstract

The human brain is undoubtedly modular, and there are numerous reasons why it might have evolved to be that way. Rueckl, Cave & Kosslyn (1989) have shown how a clear advantage in having a modular architecture can exist in neural network models of a simplified version of the "what" and "where" vision tasks. In this paper I present a series of simulations of the evolution of such neural systems that show how the advantage *can* cause modularity to evolve. However, a careful analysis indicates that drawing reliable conclusions from such an approach is far from straightforward.

Introduction

Intuitively, given the obvious potential for disruptive interference, it seems quite reasonable that two independent tasks will be more efficiently carried out separately by two dedicated modules, rather than together by a homogeneous (fully distributed) system. Certainly there is considerable neuropsychological evidence that human brains do operate in such a modular manner (e.g. Shallice, 1988). In particular, the inference from double dissociation to modularity is one of the corner stones of cognitive neuropsychology, and over recent years double dissociation between many tasks have been established, with the implication of associated modularity.

Some early neural network models seemed to indicate that fully distributed systems could also result in double dissociation (e.g. Wood, 1978) and hence cast some doubt on the inference of modularity. Since then, the potential for double dissociation in connectionist systems with and without modularity has been well studied (e.g. Plaut, 1995; Bullinaria & Chater, 1995; Bullinaria, 1999), and the early connectionist double dissociations have been seen to be merely the result of small scale artefacts. Several later studies (e.g. Devlin, Gonnerman, Andersen & Seidenberg, 1998; Bullinaria, 1999) have shown how weak double dissociation can arise as a result of resource artifacts (e.g. Shallice, 1988, p232) in fully distributed systems, but it seems that strong double dissociation does require some form of modularity, though not necessarily in the strong (hard-wired, innate and informationally encapsulated) sense of Fodor (1983). Plaut (1995), for example, has shown that double dissociation can result from damage to different parts of a single neural network, and Shallice (1988, p249) lists a number of systems that could result in double dissociation without modularity in the conventional sense. In this paper, I am not so much interested in showing how double dissociation can arise in connectionist systems without modularity, but rather, how modularity can arise in connectionist systems and hence have the potential for exhibiting double dissociation.

Of particular interest to us here is the discovery that visual perception involves two distinct cortical pathways (Mishkin, Ungerleider & Macko, 1983) – one running ventrally for identifying objects ("what"), and another running dorsally for determining their spatial locations ("where"). Some time ago, Rueckl, Cave & Kosslyn (1989) considered the interesting question of why "what" and "where" should be processed by separate visual systems in this way. By performing explicit simulation and analysis of a series of simplified neural network models they were able to show that modular networks were able to generate more efficient internal representations than fully distributed networks, and that they learned more easily how to perform the two tasks. The implication is that any process of evolution by natural selection would result in a modular architecture and hence answer the question of why modularity has arisen.

Now, eleven years later, the power of modern computer technology has finally reached a level whereby the relevant explicit evolutionary simulations are now feasible. Already Di Ferdinando, Calabretta & Parisi (2001) have established that modularity *can* evolve. In this paper, I present the results of further simulations and conclude that, whilst modularity may arise, the situation is not quite as straight-forward as the original computational investigation of Rueckl et al. (1989) suggested.

Learning Multiple Tasks

Nowadays, the basic structure of simple feed-forward neural network models is well known. We typically use a three layer network of simplified neurons. The input layer activations represent the system's input (e.g. a simplified retinal image). These activations are passed via weighted connections to the hidden layer where each unit sums its inputs and passes the result through some form of squashing function (e.g. a sigmoid) to produce its own activation level. Finally, these activations are passed by a second layer of weighted connections to the output layer where they are again summed and squashed to produce the output activations (e.g. representations of "what" and "where"). The connection weights are typically learnt by some form of gradient descent training algorithm whereby the weights are iteratively adjusted so that the network produces increasingly accurate outputs for each input in a set of training data.

In this context, the question of modularity relates to the connectivity between the network's hidden and

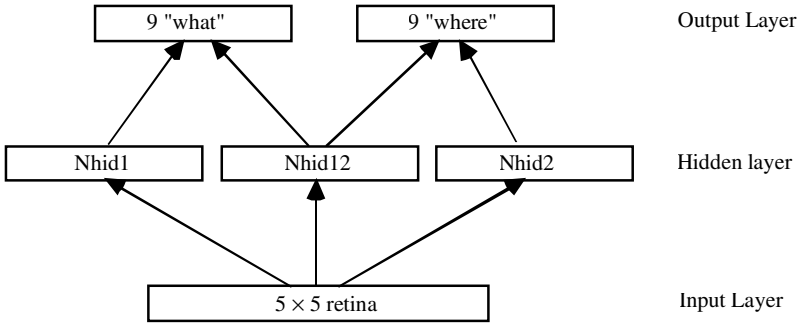


Figure 1: Architecture of the basic neural network model for the "what" and "where" tasks.

output layers. During training, a hidden unit that is being used to process information for two or more output units is likely to receive conflicting weight update contributions for the weights feeding into it, with a consequent degradation of performance relative to a network that has a separate set of hidden units for each output unit (Plaut & Hinton, 1987). However, such an extreme version of modularity with a set of hidden units (or module) for each output unit is likely to be rather inefficient in terms of computational resources, and an efficient learning algorithm should be able to deal appropriately with the conflicting weight update signals anyway. Nevertheless, splitting the hidden units up into disjoint sets corresponding to distinct output tasks, may be an efficient option. Indeed, it is hard to imagine how it could be optimal to expect a single set of hidden units to form more than one distinct internal representation.

It is well known that, when one trains a neural network using standard gradient descent type learning algorithms, the processing at the hidden layer tends to become fully distributed – in other words, there is no spontaneous emergence of modularity (e.g. Plaut, 1995; Bullinaria, 1997). However, the human brain is somewhat more sophisticated than a simple feed-forward network learning by gradient descent, and Jacobs, Jordan & Barto (1991) have shown explicitly how it is possible to set up gated mixtures of expert networks that can learn to process two tasks in a modular fashion. Such systems appear to have advantages in terms of learning speed, minimizing cross-talk (i.e. spatial interference), minimizing forgetting (i.e. temporal interference), and generalization. In a further computational study, Jacobs & Jordan (1992) have shown how a simple bias towards short range neural connectivity can also lead to the learning of modular architectures.

In this paper, I am more interested in the *evolution* of modularity than the *learning* of modularity. The old Nature-Nurture debate has come a long way in recent years (e.g. Elman et al., 1996), but it is still important to understand which characteristics are innate and which need to be learnt during ones lifetime. Moreover, as computer technology becomes more powerful, we are able to explore these issues by increasingly realistic simulations. Old ideas about the interaction of learning and evolution (e.g. Baldwin, 1896) can now be confirmed explicitly (e.g. Hinton & Nowlan, 1987). In suitably simplified systems, we have been able to observe the genetic assimilation of learnt characteristics without Lamarckian inheritance, see how appropriate innate values for network parameters and learning rates can evolve, understand how individual differences across evolved populations are constrained, and so on (e.g. Bullinaria, 2001). In the remainder of this paper I shall consider the evolution of modularity in neural network models of the "what" and "where" tasks previously studied by Rueckl et al. (1989). The lessons we learn here will be applicable to the learning and evolution of modularity more generally.

The "What" and "Where" Model

To avoid the need to repeat the extensive analyses of the learnt internal representations carried out by Rueckl et al. (1989), I shall study exactly the same simplified neural network model that they used, and explore whether the advantages of modularity they observed are sufficient to drive the evolution of modularity. I shall also follow Rueckl et al. (1989) and Jacobs et al. (1991) in emphasizing that the tasks we are simulating are vast over-simplifications of what real biological visual systems have to cope with. It makes sense to use them, however, despite their obvious unrealistic features, since they allow us to illustrate the relevant factors with simulations we can perform on current computational hardware in a reasonable amount of time.

The task consists of mapping a simplified retinal image (a 5 × 5 binary matrix) to a simplified representation of "what" (a 9 bit binary vector with one bit 'on') and a simplified representation of "where" (another 9 bit binary vector with one bit 'on'). I use the same 9 input patterns and 9 positions as in the previous studies, giving the same 81 retinal inputs for training on. Each of the 9 patterns consist of a different set of 5 cells 'on' within a 3 × 3 sub-retina array, and the 9 positions correspond to the possible centers of a 3 × 3 array within the full 5 × 5 array.

Figure 1 shows the basic network that was originally investigated by Rueckl et al. (1989). We have 25 input units, 18 output units and 81 training examples. The arrowed lines represent full connectivity, and *Nhid1*, *Nhid12*, *Nhid2* specify how many hidden units in each block. Rueckl et al. (1989) studied in detail the fully

distributed network (*Nhid1* = *Nhid2* = 0) and the purely modular network (*Nhid12* = 0). Our characterization will allow us to explore the full continuum between these extremes. If the maximum number of hidden units *Nhid* = *Nhid1* + *Nhid12* + *Nhid2* is fixed, then we need define only two innate architecture parameters *Con1* = *Nhid1* + *Nhid12* and *Con2* = *Nhid2* + *Nhid12* corresponding to the number of hidden units connecting to each output block.

Simulating Evolution

To simulate an evolutionary process for the models discussed above, we take a whole population of individual instantiations of each model and allow them to learn, procreate and die in a manner approximating these processes in real (living) systems. The genotype of each individual will depend on the genotypes of its two parents, and contain all the appropriate innate parameters. Then, throughout its life, the individual will learn from its environment how best to adjust its weights to perform most effectively. Each individual will eventually die, perhaps after producing a number of children.

In more realistic situations, the ability of an individual to survive or reproduce will rely on a number of factors which can depend in a complicated manner on that individual's performance over a range of related tasks (food gathering, fighting, running, and so on). For the purposes of our simplified model, however, we shall consider it to be a sufficiently good approximation to assume a simple relation between our single task fitness function and the survival or procreation fitness. Whilst any monotonic relation should result in similar evolutionary trends, we often find that, in simplified simulations, the details can have a big effect on what evolves and what gets lost in the noise.

I shall follow a more natural approach to procreation, mutation and survival than many evolutionary simulations have done in the past (e.g. in Belew & Mitchell, 1996). Rather than training each member of the whole population for a fixed time and then picking the fittest to breed and form the next generation, the populations will contain competing learning individuals of all ages, each with the potential for dying or procreation at each stage. During each simulated year, each individual will learn from their own experience with the environment (i.e. set of training/testing data) and have their fitness determined. A biased random subset of the least fit individuals, together with a flat random subset of the oldest individuals, will then die. These are replaced by children, each having one parent chosen randomly from the fittest members of the population, who randomly chooses a mate from the rest of the whole population. Each child inherits characteristics from both parents such that each innate free parameter is chosen at random somewhere between the values of its parents, with sufficient noise (or mutation) that there is a reasonable possibility of the parameter falling outside the range spanned by the parents. Ultimately, the simulations might benefit from more realistic encodings of the parameters, concepts such as recessive and dominant genes, learning and procreation costs, different inheritance and mutation details, different survival and procreation criteria, more restrictive mate selection regimes, protection for young offspring, different learning algorithms and fitness functions, and so on, but for the purposes of this paper, the simplified approach outlined above seems adequate. A similar regime has already been employed successfully elsewhere (Bullinaria, 2001) to study the Baldwin effect in the evolution of adaptable control systems.

The simulated genotypes naturally include all the innate parameters needed to specify the network details, namely the architecture, the learning algorithm, the learning rates, the initial connection weights, and so on. In real biological evolution, all these parameters will be free to evolve. In simulations that are designed to explore particular issues, it makes sense to fix some of these parameters to avoid the complication of un-foreseen interactions (and also to speed up the simulations). In my earlier study of genetic assimilation and the Baldwin effect (Bullinaria, 2001), for example, it made sense to keep the architecture fixed and to allow the initial innate connection weights and learning rates to evolve. Here it is more appropriate to have each individual start with random initial connection weights and allow the architecture to evolve. Then, since the optimal learning rates will vary with the architecture, we must allow these to evolve along with the architecture.

It is clearly important to fix the evolutionary parameters appropriately according to the details of the problem and the speed and coarseness of the simulations. For example, if all individuals learn the task perfectly by the end of their first year, and we only test their performance once per year, then the advantage of those that learn in two months over those that take ten is lost and our simulated evolution will not be very realistic. Since the networks were allowed to evolve their own learning rates, this had to be controlled by restricting the number of training data presentations per year to 10 for each individual. Choosing a fixed population size of 200 was a trade-off between maintaining genetic diversity and running the simulations reasonably quickly. The death rates were set in order to produce reasonable age distributions. This meant about 5 deaths per year due to competition, and another 5 individuals over the age of 30 dying each year due to old age. The mutation parameters were chosen to speed the evolution as much as possible by maintaining genetic diversity without introducing too much noise into the process. These parameter choices led to coarser simulations than one would like, but otherwise the simulations would still be running.

Experiment 1 – The Basic Model

I began by simulating the evolution of the system as stated above. For comparison purposes, this involved fixing the learning algorithm to be that used by Rueckl et al. (1989), namely online gradient descent with momentum on the Sum Squared Error cost function *E* (Hinton, 1989). As before, the target outputs were taken

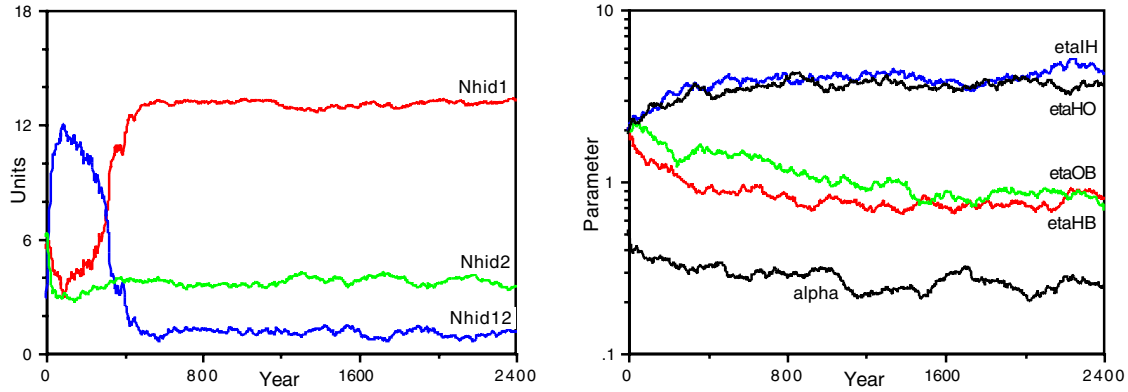


Figure 2: Evolution of the model in Figure 1 with Sum-Squared Error cost function and Log Cost fitness function.

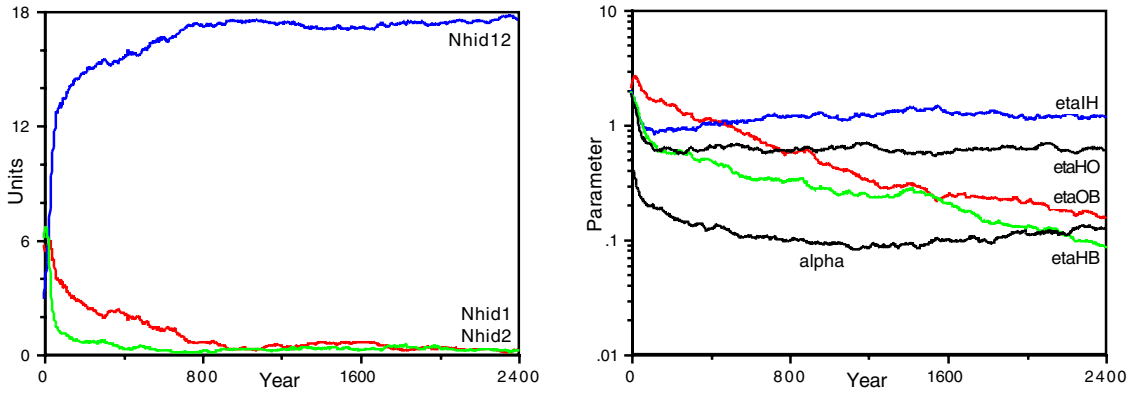


Figure 3: Evolution of the model in Figure 1 with Cross Entropy cost function and Log Error Count fitness.

to be 0.1 and 0.9, rather than 0 and 1, and appropriate outputs beyond these targets were deemed errorless. Experience indicates that the networks learn better if they have different learning rates for each of the different connection layers, and each of the different bias sets. So, to ensure that the architecture comparisons were fair in the sense that they were all learning at their full potential, each network had five learning parameters: the learning rate η_{IH} for the input to hidden layer, η_{HB} for the hidden layer biases, η_{HO} for the hidden to output layer, and η_{OB} for the output biases, and the momentum parameter α. These appear in the standard weight update equation

$$\Delta w_{ij}(n) = -\eta_L \frac{\partial E}{\partial w_{ij}} + \alpha \Delta w_{ij}(n-1).$$

Each genotype thus contained two parameters to control the network architecture, and five to control its learning rates. The Sum Squared Error cost distribution turns out to be rather skewed across the population, so the individual evolutionary fitnesses were defined to be –log(Cost).

I have found in my earlier studies (Bullinaria, 2001) that the evolution can depend on the initial conditions, i.e. on the distribution of the innate parameters across the initial population, and that the population settles into a near optimal state more quickly and reliably if it starts with a wide distribution of initial learning rates, rather than expecting the mutations to carry the system from a state in which there is little learning at all. Thus, in all the following experiments, the initial population learning rates were chosen randomly from the range [0.0. 2.0] and the momentum parameters randomly from the range [0.0, 1.0]. Following Rueckl et al. (1989), the initial weights were chosen randomly within the range [0.0,0.3].

Figure 2 shows how the innate parameters evolved when there were 18 hidden units in total (which is how many Rueckl et al., 1989, used). We see that the learning parameters soon settle down and, after a non-modular start, the population quickly evolves to take on a modular architecture with *Nhid12* near zero. This is exactly what we would expect from the Rueckl et al. (1989) study, right down the to optimal values for *Nhid1* and *Nhid2*.

Experiment 2 – Different Costs

The results of Experiment 1 make the evolution of modularity look almost inevitable. However, it would be misleading not to report on the countless simulations in which modularity did not evolve, and which could equally well correspond to human evolution, with the implication that modularity in the human brain must originate in some other manner. Figure 3 shows what can happen with one particularly reasonable alternative choice for the gradient descent cost function and

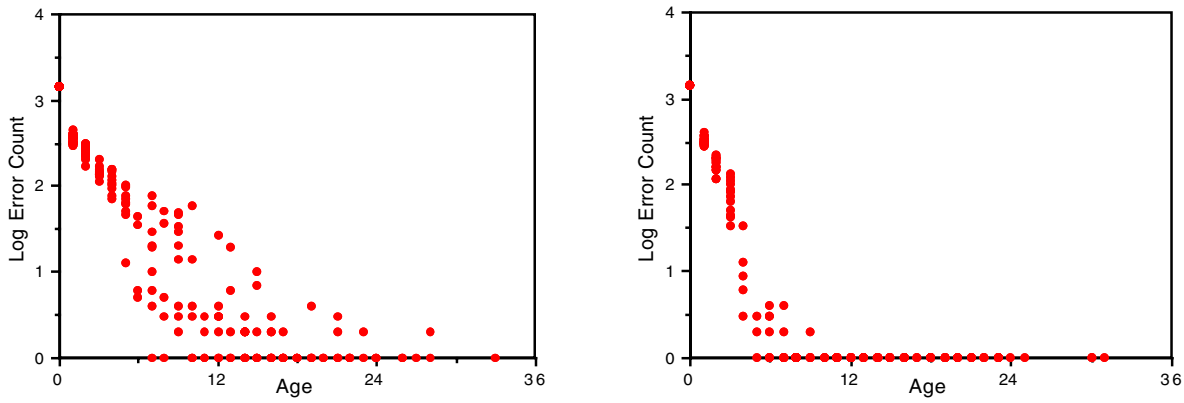


Figure 4: Comparison of evolved populations with Sum Squared Error (left) and Cross Entropy (right) cost functions.

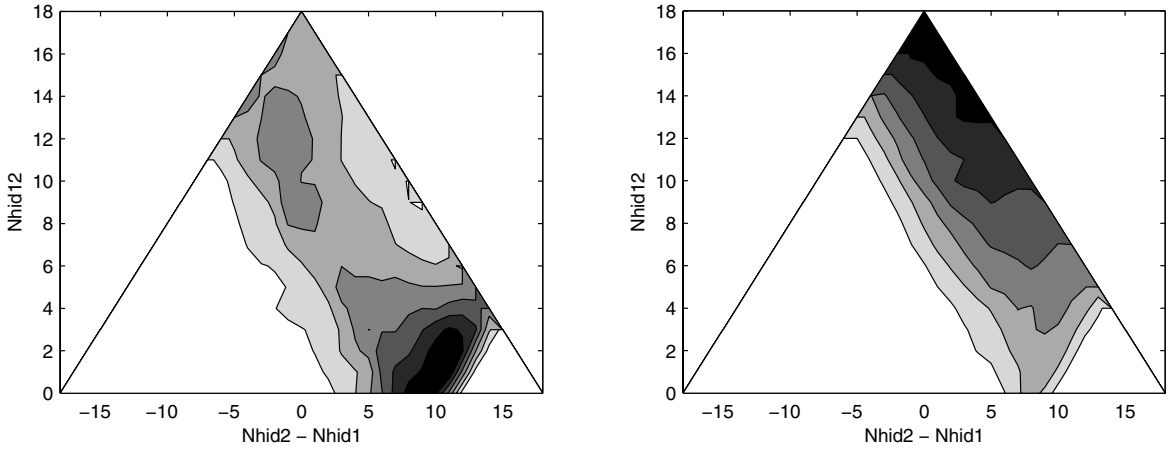


Figure 5: Mean learning times with Sum Squared Error (left) and Cross Entropy (right) cost functions.

evolutionary fitness function, namely the standard Cross-Entropy cost function (Hinton, 1989), and fitness defined by counting the total number of output units with errors above some small fixed value (0.2 say). This results in the evolution of a completely non-modular architecture. A systematic study reveals that changing the fitness between –Cost, –log(Cost), 1/Cost, –ErrorCount, and –log(1+ErrorCount) and has little effect on the results. However, the choice of cost function is crucial. Figure 4 compares the learning in the evolved populations for the Sum Squared Error and Cross Entropy cost functions with –log(1+ErrorCount) fitness. The non-modular Cross-Entropy population shows a clear superiority.

Although we should not rely on the mean learning rates to predict what will evolve (since the standard deviations, the worst and best cases, and so on, are also important), the plots in Figure 5 of the mean learning times as a function of the architecture do show quite clearly where the different optimal configurations (shown darkest) are situated.

Experiment 3 – Larger Networks

A final worry was that our simulations were suffering from small scale artefacts. Often when a network has barely enough hidden units to solve the task at hand, it behaves differently to when it has plenty of spare resources (e.g. Bullinaria & Chater, 1995; Bullinaria, 1997). Since 18 hidden units is near minimal for our task, all of the above simulations were repeated with 36 hidden units. This had little effect on the Cross Entropy simulations, but the results were rather variable with Sum Squared Error costs. Sometimes modularity evolved, sometimes it didn't, and often mixed populations arose. Apparently minor variations in the implementational details, or even just different random number seeds, could change the results completely.

Figure 6 shows the mean learning times here for comparison with those for the smaller networks in Figure 5. We see the Cross-Entropy plot has the same non-modular optimum as before, but the Sum-Squared Error case is now much noisier, with further, roughly equivalent, minima appearing in the non-modular regime. This is presumably why the evolutionary simulation results were so variable.

Conclusions

I have shown how it is possible to simulate the evolution of modularity in simple neural network models.

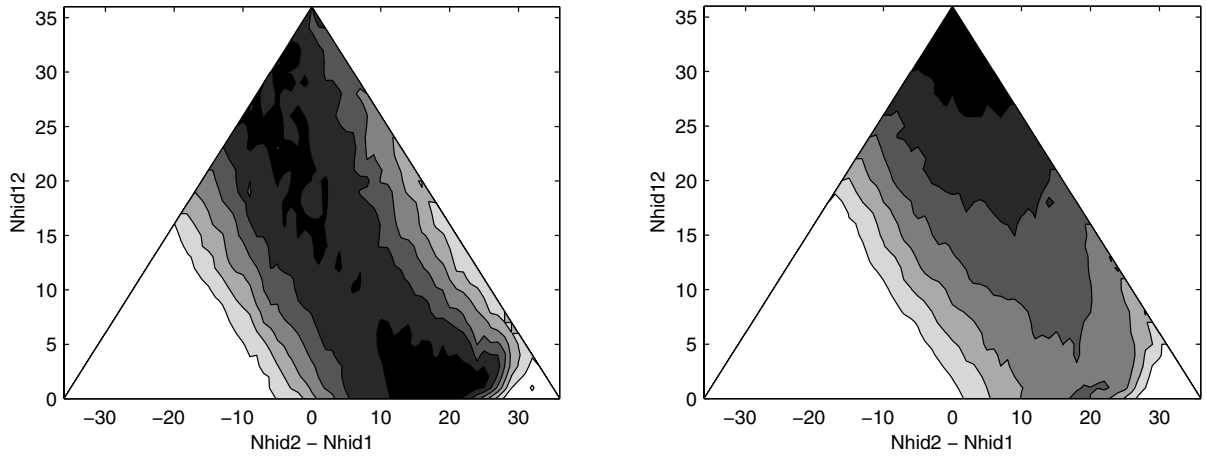


Figure 6: Large network learning times with Sum Squared Error (left) and Cross Entropy (right) cost functions.

However, drawing conclusions from them about the modularity in human brains is not so straightforward. If the results (i.e. modularity versus non-modularity) depend so crucially on such non-biologically plausible details as the learning algorithm, then it is clearly going to be rather difficult to extrapolate from them to biological systems. On one hand, we might expect that the human brain has evolved particularly efficient learning algorithms, in which case we could argue that the more efficient non-modular cross-entropy populations are the more realistic. On the other hand, real tasks are considerably harder than those used in our simulations, and so the modular populations might be deemed a more reliable representation of the actual relation between the human learning algorithm power and task complexity. The general simulation approach I have presented appears promising, but future simulations in this area will clearly have to be much more realistic if we are to draw reliable conclusions from them.

References

Baldwin, J.M. (1896). A New Factor in Evolution. *The American Naturalist*, **30**, 441-451.

Belew, R.K. & Mitchell, M. (Eds) (1996). *Adaptive Individuals in Evolving Populations*. Reading, MA: Addison-Wesley.

Bullinaria, J.A. (1997). Analysing the Internal Representations of Trained Neural Networks. In A. Browne (Ed.), *Neural Network Analysis, Architectures and Applications*, 3-26. Bristol: IOP Publishing.

Bullinaria, J.A. (1999). Connectionist Dissociations, Confounding Factors and Modularity. In D. Heinke, G.W. Humphreys & A. Olsen (Eds), *Connectionist Models in Cognitive Neuroscience,* 52-63. Springer.

Bullinaria, J.A. (2001). Exploring the Baldwin Effect in Evolving Adaptable Control Systems. In: R.F. French & J.P. Sougne (Eds), *Connectionist Models of Learning, Development and Evolution.* Springer.

Bullinaria, J.A. & Chater N. (1995). Connectionist Modelling: Implications for Cognitive Neuropsychology. *Language and Cognitive Processes*, **10**, 227-264.

Devlin, J.T., Gonnerman, L.M., Andersen, E.S. & Seidenberg, M.S. (1998). Category-Specific Semantic Deficits in Focal and Widespread Brain Damage: A Computational Account. *Journal of Cognitive Neuroscience*, **10**, 77-94.

Di Ferdinando, A., Calabretta, R, & Parisi, D. (2001). Evolving Modular Architectures for Neural Networks. In R.F. French & J.P. Sougne (Eds), *Connectionist Models of Learning, Development and Evolution.* Springer.

Elman, J.L., Bates, E.A., Johnson, M.H., Karmiloff-Smith, A., Parisi, D. & Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Fodor, J.A. (1983). *The Modularity of the Mind*. Cambridge, MA: MIT Press.

Hinton, G.E. (1989). Connectionist Learning Procedures. *Artificial Intelligence*, **40**, 185-234.

Hinton, G.E. & Nowlan, S.J. (1987). How Learning Can Guide Evolution. *Complex Systems*, **1**, 495-502.

Jacobs, R.A. & Jordan, M.I. (1992). Computational Consequences of a Bias Toward Short Connections. *Journal of Cognitive Neuroscience,* **4**, 323-336.

Jacobs, R.A., Jordan, M.I. & Barto, A.G. (1991). Task Decomposition Through Competition in Modular Connectionist Architecture: The What and Where Vision Tasks. *Cognitive Science,* **15**, 219-250.

Mishkin, M., Ungerleider, L.G. & Macko, K.A. (1983). Object Vision and Spatial Vision: Two Cortical Pathways. *Trends in Neurosciences*, **6**, 414-417.

Plaut, D.C. (1995). Double Dissociation Without Modularity: Evidence from Connectionist Neuropsychology. *Journal of Clinical and Experimental Neuropsychology*, **17**, 291-321.

Plaut, D.C. & Hinton, G.E. (1987). Learning Sets of Filters Using Back-Propagation. *Computer Speech and Language*, **2**, 35-61.

Rueckl, J.G., Cave, K.R. & Kosslyn, S.M. (1989). Why are "What" and "Where" Processed by Separate Cortical Visual Systems? A Computational Investigation. *Journal of Cognitive Neuroscience*, **1**, 171-186.

Shallice, T. (1988). *From Neuropsychology to Mental Structure*. Cambridge: Cambridge University Press.

Wood, C.C. (1978). Variations on a Theme of Lashley: Lesion Experiments on the Neural Model of Anderson, Silverstein, Ritz & Jones. *Psychological Review,* **85,** 582-591.

The Hot Hand in Basketball: Fallacy or Adaptive Thinking?

Bruce D. Burns (burnsbr@msu.edu)
Department of Psychology, Michigan State University
East Lansing, MI 48824-1117 USA

Abstract

In basketball, players believe that they should "feed the hot hand," by giving the ball to a player more often if that player has hit a number of shots in a row. However, Gilovich, Vallone & Tversky (1985) analyzed basketball players' successive shots and showed that they are independent events. Thus the hot hand seems to be a fallacy. Taking the correctness of their result as a starting point, I suggest that if one looks at the hot hand phenomena from Gigerenzer & Todd's (1999) adaptive thinking point of view, then the relevant question to ask is does belief in the hot hand lead to more scoring by a basketball team? By simulation I show that the answer to this question is yes, essentially because streaks are predictive of a player's shooting percentage. Thus belief in the hot hand may be an effective, fast and frugal heuristic for deciding how to allocate shots between member of a team.

The Hot Hand as Fallacy

Gilovich, Vallone and Tversky (1985) defined the "hot hand" in basketball as the belief that during a particular period a player's performance is significantly better than expected on the basis of a player's overall record. Gilovich et al. found that 91% of fans agreed that a player has "a better chance of making a shot after having just made his last two or three shots" and 68% said the same for free throws; 84% of fans believed that "it was important to pass the ball to someone who has just made several (two, three, or four) shots in a row." Numerical estimates reflected the same belief in streak shooting, and most players on a professional team endorsed the same beliefs. Thus belief in the hot hand appears to be widespread, and Gilovich et al. suggest that may it affect the selection of which player is given the chance to take the next shot in a game. This implication is captured by a phrase heard in basketball commentary: "feed the hot hand."

To test if the phenomena described by the hot hand actually exist, Gilovich et al. (1985) analyzed a professional basketball team's shooting over a season in order to see if streaks occur more often than expected by chance. They found that for each individual player, the proportion of shots hit was unrelated to how many previous shots in a row he had either hit or missed. Analysis also showed that the number of runs of hits or misses for each player was not significantly different from the expected number of runs calculated from a player's overall shooting percentage and assuming that all shots were independent of each other. The same independence was found for free-throws, as the probability of hitting a free-throw was the same after a hit as after a miss for a group of professional players. A controlled study of college players found the same independence between shots and found that observers could not predict which shots would be hit or missed. Thus the hot hand phenomenon appears to be a fallacy.

Why do fans and players believe in the hot hand if the empirical evidence shows that successive shots are independent? Gilovich et al. (1985) suggest that the persistence may be due to memory biases (streaks are more memorable) and misperception of chance, such as belief that small as well as large sequences are representative of their generating process (Tversky & Kahneman, 1974). Falk and Konold (1997) see the hot hand as a cognitive illusion that is another example of people's inability to perceive randomness. Again we see people inventing superfluous explanations because they perceive patterns in random phenomena.

Gilovich et al.'s (1985) result has been cited over 100 times in journals. Many of these citations are in the decision making literature, but it is also widely cited across a variety of fields. There are many citations in sports science (Vergin, 2000) and economics (Pressman, 1998), but it has also been cited in literature on law (Hanson & Kysar, 1999) and religion (Chaves & Montgomery, 1997).

There have been some challenges to Gilovich et al.'s (1985) conclusion that there are no more streaks than expected by chance in basketball, or at least to the finding's generalizability. Gilden and Wilson (1995) found some evidence of more streaks than expected in golf putting and darts, although they explain this as due to fluctuations in performance producing more streaks than expected rather than a real dependence between events. Miyoshi (2000) used simulations to suggest that Gilovich et al.'s analysis may not have been sensitive enough to detect the hot hand if hot-hand periods are relatively infrequent. However, in this paper I will assume Gilovich et al.'s (1985) conclusion that successive shots in basketball are independent events, in fact, my analysis will depend on it.

One reason for the wide interest in Gilovich et al.'s result may be the implications it appears to have for behavior. As Gilovich et al. (p. 313) state "...the belief in the 'hot hand' is not just erroneous, it could also be costly." This is because it may affect the way shots are allocated between members of a team. However, I will argue in this paper that this implication does not

necessarily follow from their analysis, rather belief in the hot hand may actually be adaptive.

The Hot Hand as Adaptive Thinking

Gigerenzer and Todd (1999) emphasize that humans and animals make decisions about their world with limited time, knowledge, and computational power. So they propose that much of human reasoning uses an adaptive tool-box containing fast and frugal heuristics that make inferences with limited time, knowledge and computation. Their viewpoint is based on a conception of bounded rationality. They contrast this with an assumption of unbounded rationality, which leads to a focus on the question: what is the normatively correct answer? Gigerenzer and Todd instead argue that one should ask: what is adaptive? That is, what behavior will meet the person's goals and uses a process that stays within the bounds of their resources?

From the point of view of basketball, whether successive shots are independent may not be the most relevant question to ask. What is adaptive for them is to maximize the number of points their team scores, so the question to be asked is does belief in the hot hand lead to more scoring than would disbelief in the hot hand?

The practical effect of belief in the hot hand is that it affects distribution of the ball. This is reflected in the statement that Gilovich et al. (1985) presented to fans and players, "it is important to pass the ball to someone who has just made several (two, or three, or four) shots in a row." Who should take the next shot is a question faced by the members of a team every time they have possession of the ball. In the absence of a time-out, it is a decision that each member of the team have to make by himself or herself in, at most, 30 seconds. Every player on a professional team is probably aware of the shooting percentage (i.e., what percentage of a players total number of shots a player hits) for each member of the team. However, knowing that one player has a 55% and another a 60% shooting percentage, does not tell one how often to give the ball to each player, given that one cannot simply give the ball to the player with the higher shooting percentage every time. Players are unlikely to be able to do a calculation to determine the optimal distribution, so fast and frugal heuristics for deciding who should take the next shot are likely to be exploited if they are effective in increasing scoring.

I propose that belief in the hot hand is such a heuristic. The basic argument is straight forward: if one accepts Gilovich et al.'s (1985) finding that successive shots are independent events, then the higher a player's shooting percentage is, the larger the number of runs of hits a player will have. Therefore, a bias to give the ball to players with the hot hand is equivalent to a bias to give the ball to players with higher shooting percentages. Giving the ball to the player with the hot hand requires no calculation, it requires only remembering the most recent shots, and it can be decided fast. Thus belief in the hot hand could be an example of adaptive thinking.

I will support this analysis with computer simulations testing whether a team that believes in the hot hand will outscore one that does not. However, I will first show empirically that players with higher shooting percentages experience more runs of hits.

Empirical Analysis

Gilovich et al. (1985, Table 1) presented the probabilities of players making a shot after runs of hits or misses of length one, two and three, as well as the frequencies of each run for players. The statistics came from analysis of the 48 home games of the Philadelphia 76ers during the 1980-81 season. In Table 1, I have re-analyzed this data to calculate for each player the proportions of his total number of shots (excluding a player's first shot in a game) which were parts of runs of hits or misses of length 1, 2, and 3.

Table 1: Proportions of players total shots that were parts of runs of 1, 2, or 3 hits, or 1, 2, or 3 misses. Correlations are between these proportions and the players' season long shooting percentage (all significant at $p < .01$)

Player	Shooting percentage	Total shots	3 misses	2 misses	1 miss	1 hit	2 hits	3 hits
Lionel Hollins	.46	371	.11	.25	.54	.46	.18	.07
Andrew Toney	.46	406	.08	.22	.53	.47	.19	.07
Caldwell Jones	.47	225	.09	.21	.52	.48	.16	.05
Clint Richardson	.50	206	.06	.16	.49	.51	.22	.10
Julius Erving	.52	836	.11	.23	.49	.51	.25	.12
Bobby Jones	.52	310	.06	.17	.47	.53	.25	.11
Steve Mix	.54	386	.06	.17	.46	.54	.25	.09
Maurice Cheeks	.56	292	.04	.13	.43	.57	.26	.11
Daryl Dawkins	.62	358	.02	.09	.38	.62	.31	.15
Correlations with shooting percentage:			-.804	-.874	-.993	.993	.954	.899

If shots are independent events, then the higher a player's shooting percentage, the larger the number of runs of hits he should have and the fewer number of runs of misses he should have. Table 1 presents the correlations between a player's shooting percentage and the proportion of his shots that were parts of runs of each length. As can be seen, all runs of misses are highly negatively correlated with shooting percentage, and all runs of hits are highly positively correlated with shooting percentage (all $p < .01$). This supports Gilovich et al.'s (1985) argument that successive shots are independent events, and also the consequence of my argument that runs are predictive of a player's shooting percentage.

Design of the Computer Simulations

In creating the simulations I strove to make them as transparent as possible and to utilize as few free parameters as possible. The basic simulation of basketball shooting had two parameters for each player: an allocation and a shooting percentage. The *allocation* is the probability of a player being given the next shot, whereas the *shooting percentage* is how often the player hits a shot. The sum of allocation parameters for all players must be 1.0 and was used to represent some underlying bias to give the ball to a player. No assumption was made regarding the source of these biases, but it was fixed for the length of a simulation. Shooting percentage reflects a player's ability to hit shots, and was also fixed for the length of a simulation, as it was for Gilovich et al.'s (1985) analysis.

The program simulated basketball shooting, with one shot per trial. On each trial a player was randomly selected to be given the next shot, with each player having the probability of being given the shot indicated by their allocation. The player given the shot then randomly hits or misses the shot with a probability indicated by the player's shooting percentage.

To simplify the simulations, rather than represent all five members of a normal basketball team, only two will be included. This reduces the number of parameters and how many players are represented should not matter with regard to the conclusions I wish to draw from the simulations. To further reduce the number of free parameters the allocation and shooting percentages were only varied for one player, and the other player's parameters were simply one minus each of the first player's parameters. Thus a whole simulation was described by just two free parameters, allowing the entire parameter space to be explored.

To simulate belief in the hot hand, a simple rule was used that determine who should be given the next shot:

1) Give the next shot to a player which has the longest run of hits (in effect, the one who hits its most recent shot), then keep giving it to that player until a miss.

2) If both players have missed their last shot, then the allocations parameters were used to select a shooter randomly.

The hot-hand could be simulated in other ways, but this seems a simple, easily understandable version, and it is parameter-less. More complicated ways of calculating who has the hot-hand would involve arbitrary parameters but produce the same pattern of results.

To test the effect of belief in the hot-hand, two simulations were run with each combination of the two parameters. Simulations with a given combination of parameters were run in pairs. In one run, the hot hand rule was turned on, and in the other it was turned off so the player to take the shot was always determined randomly using the allocation parameter. All parameter combination for allocation values from 0.01 to 0.99 were run in increments of 0.01, and shooting percentage values from 0.50 to 0.99 in increments of 0.01 (0.00 to 0.49 would simply repeat the other combinations). Thus 4851 pairs of simulations were run.

Results of the Simulations

Each combination of parameters was run for 1,000,000 trials with the hot-hand rule, and 1,000,000 without it. Each simulation produced a score, which was how many of the trials were hits. To determine the effect of belief in the hot-hand, for each parameter pair the score for the simulation without the hot-hand rule was subtracted from the score from the simulation with the hot-hand. This difference was divided by the total number of trials to yield an advantage score (adv):

$$adv = \frac{\text{(score with hot hand) - (score without hot hand)}}{\text{total number of trials}}$$

Figure 1 presents a contour graph for the 4852 (49x99) pairs of simulations (the 0.50 shooting percent parameter is excluded because when there is no difference between players, there is nothing for the hot hand to exploit). This graph represents three dimensions of information: the allocation percentage, the shooting percentage, and the adv score in favor of the hot hand simulation for that combination of parameters. The numbered contours define boundaries in the distribution of adv scores found for parameter pairs. So, for example, for every combination of parameters above the line labeled "0.2" the hot hand simulation scores at least 0.2 points per trial of the simulation (on every trial the players score 0 or 1). The areas at the bottom of the graph labeled "0.0" indicate regions in which the hot hand lost in this set of simulations. (To creates these plots I used Sigma graph, which tried to "smooth" contours resulting in these odd shapes.)

Almost all the pairs of parameters yielded positive advantage scores. Not surprisingly, the greatest advantage occurs when shooting percentage is high and allocations are low, as in effect the hot hand rule increases the allocation for the player with the higher shooting percentage. As Figure 1 shows, only when the shooting percentage is low, which is when the two players differ little in shooting percentage, does the hot hand sometimes lose. Figure 1 also shows that there is no systematic relationship between allocation and shooting percentage which results in the hot hand doing worse. When the hot hand does worse is essentially random and only occurs when there is little difference between the players for the hot hand to exploit. The most negative advantage score obtained was -0.0018 points per trial.

Table 2 shows for how many of the simulations with each shooting percentage (99 as the allocation varies from .01 to .99) the hot hand wins. When shooting percentage is equal to .50 then it should be random which simulation wins because the hot hand cannot help the better shooter when the two players do not differ. As the shooting percentage increases, and thus the difference between players increases, loses by the hot hand simulation become rarer. There were no losses by the hot hand for shooting percentages in excess of .60. There was no systematic relationship between allocation parameters and loses by the hot hand, except for a few extra loses at allocations of 0.99.

Table 2: The number of allocation values (99) that the hot hand simulation wins or loses for each shooting percentage parameter less than .61.

Shooting percentage	Hot hand loses	Hot hand wins
.50	44	55
.51	48	51
.52	21	78
.53	10	89
.54	2	97
.55	2	97
.56	1	98
.57	2	97
.58	2	97
.59	1	98
.60	1	98

.

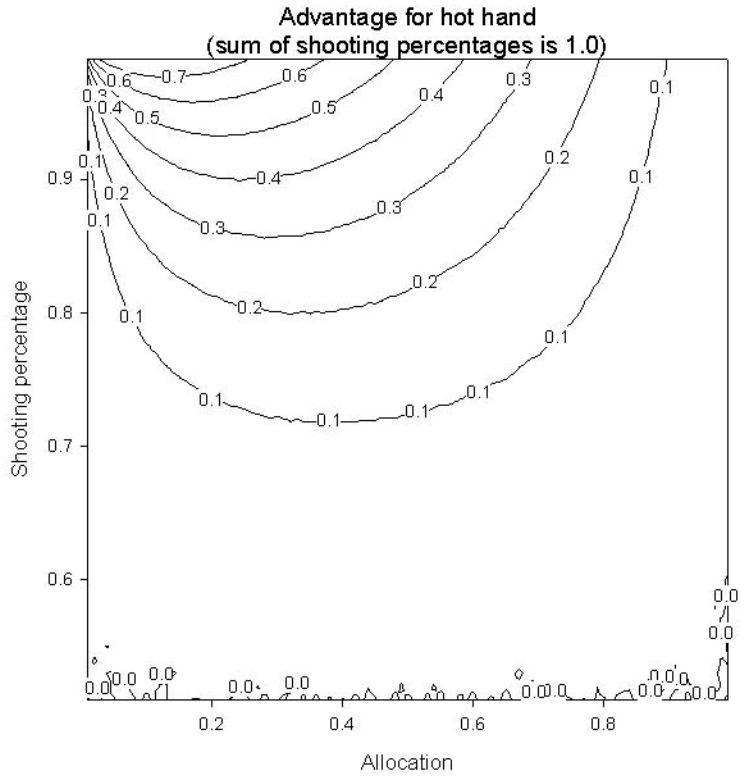


Figure 1: Contour graph showing how many points per trial the simulation using the hot hand rule came out ahead of the simulation not using the hot hand rule, for every pair of shooting percentage and allocation parameters. The lowest shooting percentage is 0.51 and highest is 0.99. The range for allocation parameters were 0.01 to 0.99.

One question that may be raised is whether the results are due to one player having a shooting percentage below 50% and the other above 50%. To check this, a set of simulations were run which were identical to the above set, but the sum of the two shooting percentages for the two players equaled 0.50. So the shooting percentage parameter was varied from 0.25 to 0.49. Very similar results were obtained with the hot hand only recording loses when the two shooting percentages were very close. Simulations of defenders acting on belief in the hot has also shown that it helps them.

Why the hot hand wins is quite clear. The hot hand rule has no effect on the percentage of shots hit by a player and does not introduce any dependencies between successive shots by the same player, instead it leads to the player with the higher shooting percentage being given more shots. In effect, the hot hand increases the allocation parameter for the player with the higher shooting percentage.

These simulations are of course not complete simulations of basketball games. Realistic simulations of ten players interacting on a basketball court are not possible, and thus this simulation cannot be compared directly to real data from basketball players. This however is not the point of the simulations, instead they are intended as an instantiation of the thought experiment conducted above, that belief in the hot hand should increase scoring by a team if successive shots are independent events. The belief increases the likelihood that the best scorers will take more shots, thus it is adaptive. The simulations are actually unnecessary if one already accepted the basic argument. However if one accepts this argument then it changes the interpretation of the data of Gilovich et al. (1985) and any other belief regarding streaks when the events making up the streaks are independent.

Fallacy and Adaptation

Is the hot hand a fallacy or adaptive thinking? In a sense it can viewed as both, it depends on what question one thinks is the most relevant one to ask.

The question that Gilovich et al. (1985) sought to answer was whether basketball players produce more streaks of hits or misses than expected by chance given their underlying shooting percentage. Their analysis showed that the answer to this question was "no", for an individual. The analysis presented here in no way challengers this result, in fact it is built into the simulations as an assumption. It may seem obvious that if a belief in the hot hand as defined for an individual is erroneous, then it must also be erroneous when applied to a team. Gilovich et al. (1985) make this quite understandable connection without comment, and thus make statements about the supposedly negative consequences for a team of passing the ball to the player with the hot hand.

The ease with which one can slip from referring to an individual's behavior to a team's behavior is reflected in the statements Gilovich et al. (1985) asked basketball fans and players to consider. Gilovich regard both of the following two statements as indicators of an erroneous belief in the hot hand:

1) "[a player] has a better chance of making a shot after having just made his last two or three shots than he does after having missed his last two or three shots"

2) "it is important to pass the ball to someone who has just made several (two, three, or four) shots in a row"

Statement 1 refers to an individual's streaks, and Gilovich et al. (1985) show empirically that it is incorrect. However, Statement 2 is about a team's decisions about how to allocate shots between players. Gilovich et al.'s data does not address this question, but the arguments and simulations presented here show that Statement 2 is correct. It is adaptive rather than an erroneous belief. From this conclusion, it is interesting to note that Statement 2 was the only statement given to the professional players by Gilovich et al. that was endorsed by every one of them.

The alternative question regarding the hot-hand is suggested by Gigerenzer and Todd's (1999) approach: is belief in the hot hand adaptive? Whether there actually are streaks in individual players' shooting is irrelevant from this point of view. The basketball players' primary goal when his or her team has possession of the ball is to maximize the number of points that the team scores (notwithstanding the behavior of some current NBA stars), as that is what determines the outcome of the game. If belief in the hot hand (as defined as giving the ball to the player experiencing a streak) tends to increase point scoring as compared to when the hot hand is disregarded, then the hot hand is adaptive thinking rather than a fallacy. "Feed the hot hand" can be seen as a fast and frugal heuristic for making a critical decision: who should get the next shot? Belief in the hot hand provides a quick and easily assessable answer to this question. (This is not to imply that the hot hand is the only way, or always the best way, to make this decision. Like any heuristic, it may fail.)

If there were fluctuations in a player's underlying shooting percentage, which could arise for various reasons, then the hot hand provides a further advantage over any calculations based on a shooting percentage or some other product of the history of a player. The hot hand is immediately sensitive to fluctuations because if a player's shooting percentage changes then his or her expected number of streaks will be affected immediately. The impact of fluctuations on a player's season long shooting percentage, or any other statistics,

will be delayed. Gilden and Wilson (1995) argue that such fluctuation could create streaks despite independence between successive events. Whether Gilovich et al.'s analysis was sensitive enough to detect such streaks is a question raised by the analysis of Miyoshi (2000) who points out that it would depend on the frequency and duration of such events. However, even if there are no fluctuation driven streaks in basketball, there may be in other multi-trial tasks, and thus belief in the hot-hand may be a general heuristic that people learn is effective in a variety of situations.

It could be argued that even if the belief in the hot hand is adaptive then it may originate and be sustained by a fallacy regarding the streaks of individuals. Thus basketball players may have just got lucky that their fallacy helps rather than hinders them. I have presented no evidence regarding the origin of the belief in the hot hand, and I doubt that players are consciously using the analysis I present here to support their belief in giving the ball to the player with the hot hand. However, it could be argued that what sustains belief in the hot hand is simply that players have learned that giving the ball to the player experiencing streaks has a positive outcome. The work on implicit learning shows that people may not necessarily know what they have learned. Nisbett and Wilson (1977) review evidence that people may make appropriate decisions without conscious awareness of the true reasons why they made that decision, and then they may make up plausible sounding reasons for their behavior. The erroneous beliefs fans and players make about the consequences of streaks by individual players may simply be an attempt to rationalize a behavior they have learned is adaptive. Thus belief in streaks for individuals may be a misanalysis of the reasons for an accurate perception of the hot hand as it applies to a team play, rather than a misperception of sequences which appears to be the basis of the gambler's fallacy. The connection between the gambler's fallacy and the hot hand, which may be related but describe the opposite behavior (i.e., go with the streak, verse go against the streak), may be a fruitful area for future research.

In summary, the final answer to the question posed by the title depends on which question one prefers to ask. Either, what is a normatively correct way of describing the performance of an individual basketball player, or what may lead to a higher score in a game? Even though the relevance of both questions can be seen, to an adaptive organism the later question should be more important on the basketball court.

Acknowledgments

I would like to thank Tom Carr, Bryan Corpus, Patrick Gallagher, and Mareike Wieth, for ideas and comments on earlier drafts of this paper.

References

Chaves, M., & Montgomery, J. D. (1996). Rationality and the framing of religious choices. *Journal for the Scientific Study of Religion, 35*, 128-144.

Falk, R., & Konold, C. (1997). Making sense of randomness implicit encoding as a basis for judgment. *Psychological Review, 104*, 301-318.

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive tool box. In G. Gigerenzer, P. Todd, & the ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 3-34). New York: Oxford University Press.

Gilden, D. L., & Wilson, S. G. (1995). Streaks in skilled performance. *Psychonomic Bulletin & Review, 2*, 260-265.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology, 17*, 295-314.

Hanson J. D., & Kysar D. A. (1999). Taking behavioralism seriously: The problem of market manipulation. *New York University Law Review, 74*, 630-749.

Miyoshi, H. (2000). Is the "hot-hands" phenomenon a misperception of random events? *Japanese Psychological Research. 42*, 128-133.

Pressman, S. (1998). On financial frauds and their causes: Investor overconfidence. *American Journal of economics and sociology, 57*, 405-421.

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Vergin, R. (2000). Winning streaks in sports and the misperception of momentum. *Journal of Sport Behavior, 23*, 181-197

Nisbett, R. E., & Wilson, T. D. (1978). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 213-259.

Modelling Policies for Collaboration

Mark Burton (mburton@arm.com)
ARM; 110 Fulbourn Rd
Cambridge CB1 9NJ, UK

Paul Brna (paul@cbl.leeds.ac.uk)
Computer Based Learning Unit; Leeds University
Leeds LS2 9JT, UK

Abstract

There are different ways in which learners can be organised when working together collaboratively. The issue discussed here is how to organise collaborative activities to help students *learn* how to collaborate — in particular, which *policy* is the best one to adopt for managing group communication to achieve this end. Here, different policies are considered which might apply during collaborative task performance. A computation model (Clarissa) is used to examine how these different policies perform in a collaborative task performance. The behaviour of the model is examined from the point of view that participants need to experience a range of dialogue roles which reflect the linguistic functions of utterances that are involved in collaboration. The results derived from the analysis of the behaviour of the computation model complements empirical work on small group behaviour.

Models of Collaboration

Collaboration may be mainly "in the eye of the beholder": in many educational settings the children collaborating are scarcely aware of any rules governing their work together. This leads to the notion that there are degrees of collaboration, that there may be a value for methods of assessing *how much* collaboration there is.

When the process of collaboration is examined, there is an issue about whether the current task is divided into parts tackled by different collaborators or whether collaboration entails synchronous effort with no division of the task. This task centred perspective may be appropriate when the focus is on task completion but other issues become important if the emphasis is on learning outcomes, or on the process of collaboration itself.

Whether collaboration is the means to the end of learning some domain material or whether collaboration is in some sense the end itself. In the former case, learning efficiency may be most important but in the latter case, efficiency is not the key issue — indeed, when learning to collaborate is the primary goal it may well be that inefficient domain learning is inevitable.

One view of collaboration is that part of the process of task performance requires that there is some agreement about the nature of the task — and this can be quite complex. For example, the view of Roschelle and Teasley that collaboration is "a coordinated, synchronous activity that is the result of a continued attempt to construct and maintain a shared conception of a problem" (Roschelle and Teasley, 1995). For Roschelle and Teasley, part of task performance involves synchronous effort with no division of that particular part of the task. This approach may serve to define when collaboration is taking place and when it is not — but it provides little help in determining, for example, the quality of the collaboration or how learners become more proficient at collaboration.

In contrast, here, learners are assumed to have some task to perform together, and that some way is required of estimating the quality of any perceived collaboration. The emphasis is therefore on the collaborative process as opposed to the collaborative state. The main issue that is examined is learning to collaborate so domain-based learning outcomes are of less interest.

The key notion is that good collaboration is characterised by an even distribution of dialogue roles. The justification is based on a model of distributed cognition in which cognitive processes communicate with each other (Minsky, 1987). The further assumption is that learning to collaborate is best for the group of learners when each learner exercises all the dialogue roles available, and all dialogue roles are exercised equally.

Defining dialogue features as dialogue roles, and analysing dialogue for these roles gives a broader framework in which to view collaboration. This is useful since Webb and Cohen (and others) come up with different necessary conditions for good collaboration (Webb, 1983; Cohen, 1994). For example, while Cohen suggests that the task may be responsible for determining what roles are possible in a situation, she also argues that roles (and dialogue features) are crucial in themselves.

Policies for Collaborative Activity

The aim is to examine which policy is likely to be "best" in terms of some measure. To achieve this, a number of likely candidates are described. The main hypothesis investigated is that "Normal" collaboration, unrestricted by any enforced distribution of dialogue roles[1], except that certain social norms operate, does not yield the optimum results for learning to collaborate (according to the measure we adopt below).

The basic policy that is closest to perhaps the most familiar form of collaboration is termed *Free*. For this

[1] i.e. linguistic functions of utterances used in collaboration

"normal" case, agents are permitted to drop roles whenever they wish. The primary restraint on their role usage is that two agents are not permitted to have the same role at the same time — roughly, normal social behaviour.

Other forms include *Multi*, *Swap* and *Polite*. Viewed as constraints on behaviour, some of these forms can be combined to yield, for example, *MultiSwap* and *MultiPolite*. The *Multi* policy allows agents to use the same role at the same time (to do so, the number of roles had to be slightly simplified). The resulting conversations will be "strange" in that the dialogue will seem to lose its coherence to some degree as questions are asked together, and agents may "reply" with other questions (Clarissa agents have a very good memory which is a key feature of this collaborative situation).

For the *Swap* policy, the opening of a new dialogue game is an indication of a new episode, and at this point, agents drop the roles they are playing (it does not follow that they would choose different ones for the next episode). This approach is derived from an observation by Soller who found that in the cases where people do choose different roles at the beginning of new episodes, collaborative activity is more beneficial (according to her measures) (Soller, 1997). The *Polite* policy arranges for agents to 'drop' their roles at the *end* of episodes. Rather than swapping roles at the beginning of a new episode, the participant who has lead an episode stands back for the next. In other words, if you have been taking the lead for a while, stop, and let somebody else take the floor.

MultiSwap, in an educational context, is equivalent to a situation in which collaborators can ask all the questions they have, and say everything they can about a problem. They pay careful attention to everything said by their partners. Questions, comments and suggestions must be noted down. An alternative approach is *MultiPolite* collaboration which differs from *MultiSwap* in the same way that *Polite* and *Swap* differ.

Clarissa's Architecture

The architecture is divided into two main subsystems: the cognitive and dialogue systems. The cognitive system is broken into separate units (cognitive processes), which then communicate with each other.

This architecture is rooted in notions of distributed cognition. Hutchins draws a parallel between the division of labour and the division of cognitive processes (Hutchins, 1995). Collaborative activity may not be obviously divided at the domain level, but individuals may each be exercising different cognitive processes. Clark and Chalmers advocate that cognition is not bounded by the skin, but encompasses a wider context (Clark and Chalmers, 1998). They develop active externalism, noting that internal cognitive processes are coupled in a two-way interaction with external entities. Bearing in mind that Vygotsky talks in terms of dialogue behaving as stimuli in the process(es) of solving a task, Clark and Chalmers go on to discuss the function of language as the means of coupling cognitive processes to the external environment. They conclude that the ability to communicate between internal and external processes, extended cognition "...is no simple add-on extra, but a core cognitive process". Here, this is taken as meaning that the mechanism by which processes are coupled (and communicated) internally are very similar to those which allow external coupling.

All communication that might normally be regarded as occurring within the cognitive system are made using the dialogue system (see below). The decision about which utterances are 'hearable', and which passed directly between cognitive processes within the agent are made within the dialogue system. Other than with respect to communication, individual cognitive processes may be implemented in any way. Clarissa itself uses a very simple forward chaining production rule system.

The dialogue system uses a 'dialogue game' mechanism to maintain the focus[2], and form a prediction about what might be said next. A dialogue game is defined to be a state machine which represents the entirety of possible dialogue utterances and the order in which they can occur. Parallel dialogue threads are used to let agents keep a number of topics active at the same time.

The dialogue role mechanism is used to control the communication between cognitive processes. A dialogue role can be seen as defining a (complex) zone within a dialogue game, a 'sub dialogue game'. Roles that one agent is playing cannot be used by another. Roles are swapped frequently, and the effect of this restriction is to model the way people normally respond to each other, one asking a question, the other replying and so on. To decide what is said next, both the role that individuals are playing, and the dialogue game is examined. Clarissa allows for a variety of dynamic mechanisms for distributing roles throughout the period of the ongoing dialogue.

The dialogue system must allow messages to be passed between the cognitive processes found in the cognitive system. To achieve this communication, a new abstraction is introduced, a *dialogue goal*. A dialogue goal is an expression of the communicative act (Maybury, 1993) that the cognitive system wishes to take place. The dialogue system can then choose how best to achieve this goal. The cognitive process initiates a dialogue goal whenever it wishes to pass a message to another process. These are 'thrown over the wall' to the dialogue system which requires that the goals the cognitive system generates are understandable by the dialogue system so there will be some dependency between these two systems. Goals so delivered, and acted upon by the dialogue system may eventually be completed. The desired result is that the relevant messages are delivered back to the cognitive system. To do so requires that there is some mechanism in the dialogue system for passing information back to the cognitive system.

Any number of Clarissa agents can collaborate, and they do so within a specific computational context. The task context used for this paper requires that the simulated novice physics students have to draw a diagram to

[2]Related to McCoy's approach (McCoy and Cheng, 1991)

represent how energy flows through a n electrical circuit with a charged battery connected to a bulb (with two wires), and makes a bulb shine. Space precludes a full discussion of the architectural issues — see (Burton, 1998; Burton et al., 2000) for more details.

Experimenting with Clarissa

The Clarissa system was set up to run two Clarissa agents collaborating on the standard test problem for a hundred runs for each of a number of policies (Clarissa agents act in a distributed network). The results were interpreted in terms of the distribution of roles throughout the task performance. This is in line with the underlying claim that exercising the range of available roles is advantageous, and that role usage is balanced between the participants.

Role usage is approximated by the number of utterances which are made by participants while an individual is 'playing' that role. Different policies are compared so that a provisional determination can be made about the ways in which collaboration can be organised to yield good collaboration. To analyse the various different collaborative environments samples of 100 pairs were used. Table 1 presents the important statistical measures.

There are two key ideas in the (non-standard) statistical analysis. Firstly, there is an analysis of the degree to which agents are split by the collaborative environment within a pair, so that one performs differently from the other. Secondly, it is also possible to discover something about how they perform differently by looking at the correlation between the agents. It might be hoped that agents in a pair perform equally. The above statistic (the significant split) will tell us whether it is likely that the differences between the participants in a pair can be accounted for simply by the variation in the population in general. Ideally, agents in a pair that uses a given role a lot shares this usage between the agents. In other words, if one agent uses a role a lot, the other should do likewise. For the purposes of this study, it is desirable to have a positive correlation between the two agents for all of the roles investigated. Negative correlations are likely to result in significantly split role usage. A negative correlation implies that the collaborative environment, of itself, is inducing the agents to divide the roles unevenly.

Any role usage which seems to be negatively correlated between two agents is of interest. Positive correlations are to be encouraged, as they imply that the agents are splitting the role between themselves sensibly. These statistics are used to determine how many of the roles are being played equally by participants in a pair, and how many are unevenly split, for a given collaborative environment. Additionally, the number of times agents receive interesting information from their partner is used[3]. In summary, the metric for good collaboration, according to the criteria that roles should be evenly distributed is that: Neither roles, nor the number of times agents receive interesting information from their partner, should be significantly split, or negatively correlated.

In Table 1 the results are interpreted in the case of the *Free* policy, 5 out of the 7 roles, and the number of "interesting" moves received by agents (6 out of 8 in other words) are significantly negatively correlated. this statistic indicates the degree to which the two agents in the pair share out the usage of a role.

In this case, as one agent used a role more, the other uses it correspondingly less, (and similarly the number of interesting events recoded by each agent follows a similar pattern). Correspondingly, all those roles which are significantly split, are also significantly negatively correlated (again, at the 1% level). Here the statistic is measuring the propensity of an agent to use a role more while their partner uses it less. Thus in the *Free* case, 6 out of 8 roles are negatively correlated, indicating that one agent tends to use the role more while the other uses it less.

The effects of this collaborative environment can be characterised in terms of the measure defined above as having 6 out of 8 roles significantly split and negatively correlated and a mean number of interesting events of 24.9 which is significantly split between agents in a pair, with the mean difference being 12.8.

The result of this experiment suggest that if all other factors are equal, students that collaborate together, with no control on their dialogue, are likely to benefit unequally from the experience. The first student who takes control of the conversation will remain in control, while the other participants adopt a more subservient role. This is undesirable from the perspective adopted here, but possibly common in small group work.

Examining the *Swap* policy, the reason role for the *Swap* policy is still significantly split to roughly the same degree as for the normal environment, but overall the picture is very different (Table 2). Now only 2 out of the 8 roles (and events) are significantly split while 5 are negatively correlated. The statistics suggest that it is no longer possible to be certain (at the 1% level) whether there is a significant split between agents in the pair. But in this case the mean of the difference between the agents is still relatively high (11.9) and fairly similar to the previous environment (12.8). The difference here is in the spread of the original population.

The correlation coefficient tells a similar story. It is worth noting at this point that in the case of the "interesting" events, 8 outliers have been removed from this data sample (by the procedure described above). If this had not been done, the resulting correlation would be about 0.614, significantly positive. However, in the overwhelming majority (92%) this would be misleading, as in their case the correlation is calculated as -0.462, significantly negatively correlated. The interest here is in the majority, so it is important to remove the outliers. The figure of -0.462 is still significant and leads us to believe that this sort of distribution would not have happened by chance, but it is clearly less so than the -0.99 seen in the previous example.

In common with the "interesting" case above, a number of the roles are split, but not significantly so (at the

[3] In this work, "interesting" information is information that leads to an agent's knowledge base changing.

Table 1: Results for the *Free* policy (**Bold** entries statistically significant)

	Interesting	check	reason	generate	response	interface	argue	question
Correlation	**-0.972**	-0.189	**-0.99**	**-0.99**	**-0.765**	**-0.985**	**-0.974**	-0.11
Split	**11.10**	1.41	**19.75**	**19.69**	**3.12**	**16.13**	**12.39**	1.28

Table 2: Results for *Swap* environment (**Bold** entries statistically significant)

	Interesting	argue	check	generate	interface	question	reason	response
Correlation	**-0.462**	**-0.56**	-0.19	**-0.58**	**-0.281**	-0.0105	**-0.594**	0.186
Split	2.22	2.39	1.35	**2.67**	1.68	1.13	**2.68**	0.96

1% level). They are on the borderline, and are rejected by our relatively stringent test. We expect that if roles are negatively correlated, they are also likely to be significantly split. In this case, while the correlation statistic yields a value which is significant at the 1% level, a number of the "split" statistics are not significant at this level. The implication is that the correlation between the two statistics is not as linear as might have been suspected. The fact that roles are significantly negatively correlated, but not significantly split means that both statistics need to be considered. The interpretation of such a scenario is that in this environment, the inclination of participants is that if one starts to use the role more, the other will use it less. However, this happens to a lesser degree than the overall distribution of the way in which participants use the roles anyway and (at this given level of significance) they do not divide the way in which they use the roles.

In conclusion, the *Swap* environment does indeed produce somewhat better collaboration according to our measure. This finding indicates that our first hypothesis is correct. Simple, unconstrained collaboration is not necessarily the most effective. But the *Swap* constraints still produce an environment in which a number of roles are split between participants, and many which are negatively correlated. In short, *Swap* is better, but not best! The mean of the number of interesting events found by the agents has also risen, and, as seen above, the degree to which the level of interest is split between partners in a group has fallen. However, the rise is less than our chosen statistical degree of significance, so the idea that this has happened by chance is not rejected.

Multi, the next policy to be examined, is based on allowing participants to put forward several ideas in parallel, not necessarily commenting on any of them. In practice, each participant may ask several questions, and their partners may not necessarily answer them directly (and likewise for other dialogue utterances).

It is important to remember at this stage that the Clarissa agents have infallible memories. If this form of collaboration were to be attempted between human participants, not only would the participants need to be trained in the dialogue structure (or forced to use it by an interface), but they would also require some assistance in terms of remembering what had been said.

In this case, neither the interesting events, nor the reason role, are significantly split (Table 3). However the interesting events show a strong negative correlation. Again, the pairwise difference being not significant in this case implies that while there is a tendency for one participant to receive more interesting events from their partner than vice versa, the degree to which this happens is less than the distribution of how interesting events are seen by participants anyway.

The only other role to be negatively correlated is also significantly split. That is the "interface" role. Effectively, this is the role used by participants wishing to use the mouse. In many respects we would guess that this would be the most difficult role to make sure was evenly divided between the participants.

The complete metric for the *Multi* policy is measured across 6 roles (including the "interesting" events). In this case, "check" has been amalgamated with "question" and "argue" with "response". These simplifications were made to keep the complexity of the dialogue game definition within a reasonable level (since all roles must be replicated for *Multi* and all replicated roles must be related to all others). The results are 1 out of 6 have significantly split role usage, and 2 have negative correlation. The mean of the interesting events is 26.6. Again, this mean is not significantly different from either of the means of the other two samples.

The collaborative environments which *Multi* represents seem to have allowed agents to distribute the roles more evenly, but at the cost of relying on agents having a perfect memory for what has been said by others, (so that they may return to these previous statements). The degree to which roles are more evenly distributed in this case could was not predicted. It seems that allowing agents to play the same role together allows them to avoid situations in which one agent simply takes control. This is a useful finding.

The model seems to be implying that it is not necessarily the case that the individuals should be encouraged to practise different collaborative skills than their partners (at any one time), but that it is both acceptable and possibly beneficial at least to allow them to practise sim-

Table 3: Results for the *Multi* policy (**Bold** entries statistically significant)

	Interesting	generate	interface	question	reason	response
Correlation	**-0.649**	-0.8	**-0.759**	0.121	0.332	0.043
Split	2.59	0.39	**3.36**	1.07	0.80	1.10

ilar skills concurrently. A rise was noticed in the interest level shown in *Swap* collaboration above, which was not significant at the 1% level. The next step is to examine a policy which exploits this rise, in conjunction with the better role usage seen for the *Multi* policy.

Both the *Multi* and the *Swap* constraints can be combined, since they are addressing different aspects of the environment. The former is about the roles which can be used together at any one time, the latter is about the points at which agents are expected to swap roles. It is not so clear that by combining them the results are positive. Not only are the results positive, taking the best from both individual schemes, but that they are better than one would have imagined (Table 4). Neither of the two data sets show any degree of split or negative correlation, and this is common across all 6 data sets, including the "interface" role which in the "multi" environment was still split. In the case of the "reason" role there is a degree of positive correlation (which is significant at the 1% level, as it is greater than 0.254). This is highly desirable as it implies that one participants uses the role more, the other will follow suit.

Furthermore, the level of interest is higher than seen in the "multi" environment. This must be discounted as the difference is very small and statistically could easily have happened by chance. Indeed the probability of the rise in the degree of interest from the *Free* case (24.9) and this case (29.0) happening by chance is still about 27%, a very long way from the 1% cut off level.

Nonetheless, this mode of collaboration is clearly much better according to our metric, especially in so far as it seems to encourage the even distribution of all dialogue roles. This implies that all the participants in the collaboration will have an opportunity to practise and observe all of the dialogue roles that are available, and by doing so, if these roles have been chosen to reflect underlying cognitive processes which are pedagogically interesting, the participants should have the opportunity to practise, observe and improve their ability to execute those underlying processes.

It has been our goal to find such a cognitive environment. Having done so it is necessary to reflect on how this environment relates to a realistic situation with human collaborators. It was previously noted that the human participants would need some additional assistance to allow them to keep track of what everybody had said. (This is common in most meetings, and often some such device, like paper and pencil, is encouraged in educational collaboration). They also need to be encouraged to speak their mind, ask all their questions, say everything they can about a problem, with, perhaps little regard for their partners. In another respect, they must of course take careful note of everything their partners say, such that they can respond at some point. This satisfies the "multi" constraints. Finally the *Swap* constraints imply that participants should keep asking questions, if that is what they start doing; keep making statements about what is known about the problem; keep proposing deductions, suggesting things to do next. Participants should continue concentrating on one "role" until the beginning of a new "dialogue episode". The beginning of a new dialogue game has been used as such a marker. The key is to guarantee that these events happen frequently enough that participants do not get frustrated and bored by being captured in one role for too long a time.

Of course it is necessary to re-emphasise that this is the finding of a simulator which is built in a cognitively plausible manner, but has necessarily simplified many issues. The results of simulated model can do no more than suggest ways in which others may take this research further, having decided upon their pedagogic requirements, and the dialogue roles that they hope will fulfill these.

An alternative approach to the *Swap* form of role swapping, as has been mentioned previously, is the *Polite* mode. Here, rather than everybody swapping roles at the beginning of dialogue episodes, participants who finish dialogue episodes agree to drop their roles, allowing somebody else to take the floor. This may, in some situations, be an easier form of dialogue to explain and encourage in participants. Simply, having led the discussion about a specific topic, a participant should allow their partner to lead the next. This behaviour can be combined with the *Multi* behaviour previously examined (*Polite* is very similar to *Swap* so it is not presented here).

The findings, shown in Table 5, for this mode of collaboration are similar to those of *MultiSwap* collaboration looked at above. None of the roles are significantly split or negatively correlated. The mean number of interesting utterances received by one partner from the other in this case has risen to 33.9 (s=8.41). The weighted standard deviation of this population and the population of *Free* collaborative partners (s=1.63) is 8.57 (2 d.p.). The difference in the mean is 9, which is just 1.05 SD. This is still not significant at the 1% level, indeed it would be expected in 29.38% of normally distributed cases which had similar means. Little can be said about the increase in the level of interest seen, except to note its increase, and to note that it is slightly above that for *MultiSwap*.

There is then very little to choose between the *MultiSwap* environment and the *MultiPolite* environment. Both yield good collaboration as measured by our metric. The decision about which to use may rest on the ease

Table 4: Results for *MultiSwap* environment (**Bold** entries statistically significant)

	Interesting	generate	interface	question	reason	response
Correlation	-0.238	0.874	-0.223	0.0662	0.632	0.0787
Split	1.53	0.30	1.45	1.00	0.47	1.11

Table 5: Results for *MultiPolite* environment (**Bold** entries statistically significant)

	Interesting	generate	interface	question	reason	response
Correlation	0.194	0.691	0.0567	0.128	0.666	0.0684
Split	0.93	0.45	1.10	1.01	0.50	0.98

with which the various different behaviours demanded by the collaborative environment can be taught to the participants or engineered into the environment in which they are collaborating.

Discussion

Clarissa agents simulate a form of collaboration. They offer a test-bench for investigating collaborative activity. Various different policies for collaboration have been, and others could be, examined with Clarissa. Different opinions about what makes for good collaboration can be used to investigate a variety of different collaborative situations. The notion that good collaboration is characterised by an even distribution of roles has been adopted and some hypotheses about collaboration have been evaluated given the starting point that even role distribution is important. The following results have been found:

- The best collaborative environment found involves participants having the ability to communicate in a non-normal way, and having some form of aide-memoire which would allow this abnormal communication to be effective. This finding was not predicted.
- Socially "normal" role usage is not conducive to educationally beneficial collaboration. One agent takes control of the conversation, and the social norms are such that the agent remains in control. This problem seems to be addressed by the multi-role distributions. Allowing participants to adopt the same roles at the same time is not socially normal, but seems to have a very significant effect on the quality of collaboration.

In the best form of collaboration that has been found, the environment involved participants swapping roles frequently. They can be instructed to do so either when they finish, or start, a new dialogue episode. The simplest instruction may be "if you have been dominating the discussion, when you come to the end of a topic, allow somebody else to take the lead". Clarissa has given interesting results that can be taken further.

Acknowledgments

Mark Burton was supported in this work through an EPSRC PhD Studentship.

References

Burton, M. (1998). *Computer Modelling of Dialogue Roles in Collaborative Learning Activities*. PhD thesis, Computer Based Learning Unit, The University of Leeds.

Burton, M., Brna, P. and Pilkington, R. (2000). Clarissa: A laboratory for the modelling of collaboration. *International Journal of Artificial Intelligence in Education*, 11(2):79–105.

Clark, A. and Chalmers, D. (1998). The extended mind. *Analysis*, 58(1).

Cohen, E.G. (1994). Restructuring the Classroom: Conditions for Productive Small Groups. *Review of Educational Research*, 64(1):1–35.

Hutchins, E. (1995). *Cognition in the Wild*. MIT Press, Cambridge, MA.

Maybury, M. (1993). Planning multimedia explanations using communicative acts. In Maybury, M., editor, *European Conference on Hypertext' 94*. MIT Press, Cambridge, MA.

McCoy, K. and Cheng, J. (1991). Focus of attention: Constraining what can be said next. In *Natural Language Generation in Artificial Intelligence*, chapter 4, pages 103–124.

Minsky, M. (1987). *The Society of Mind*. MIT Press, Cambridge, MA.

Roschelle, J. and Teasley, S. (1995). The construction of shared knowledge in collaborative problem solving. In O'Malley, C., editor, *Computer Supported Collaborative Learning*, pages 69–97. Springer-Verlag, Heidelberg.

Soller, A. (1997). Personal Communication.

Webb, N.M. (1983). Predicting Learning from Student Interaction: Defining the Interaction Variables. *Educational Psychologist*, 18(1):33–41.

Evaluating the Effects of Natural Language Generation Techniques on Reader Satisfaction

Charles B. Callaway (cbcallaw@eos.ncsu.edu)
James C. Lester (lester@csc.ncsu.edu)
The IntelliMedia Initiative
Department of Computer Science
North Carolina State University
Raleigh, NC 27695 USA

Abstract

We are witnessing the emergence of a new technology for dynamically creating stories tailored to the interests of particular readers. *Narrative prose generators* offer much promise for literacy education, but designing them for maximal effectiveness requires us to understand their effect on readers. This article describes the evaluation of STORYBOOK, an implemented narrative prose generation system that produces original fairy tales in the Little Red Riding Hood domain. STORYBOOK creates two to three pages of text consistently represented at the deep linguistic structure level. Because of this, we can formally evaluate multiple versions of a single story and be assured that the content is identical across all versions. We produced five such versions of two separate stories which were compared by a pool of twenty upper division students in English and analyzed with an ANOVA test. While the results are most informative for designers of narrative prose generators, it provides important baselines for research into natural language systems in general.

Introduction

The emerging technology of narrative prose generation, which dynamically creates stories tailored to the interests of particular readers, offers great promise for literacy education. However, to design effective narrative prose generation software for literacy education, it is important to understand how students perceive texts created by these algorithms. Do the results of studies based on human-produced texts apply? How does computer control of minute aspects of text production affect readers? Do readers have quantitative reactions to fundamental alterations in texts as we expect they would?

As a result of recent work in formally evaluated language generation technology (Smith & Hipp 1994; Robin & McKeown 1995; Allen *et al.* 1996; Callaway & Lester 1997; Lester & Porter 1997; Young 1999), we are seeing an increased awareness of the issues involved in successfully generating texts dynamically for specific target audiences. However, these systems are focused more towards task effectiveness evaluations or explanation generation and are not suitable for the significant difficulties in creating literary narratives. And while there exist story generation systems capable of producing narratives (Meehan 1976; Lebowitz 1985; Lang 1997), none of these systems has been formally evaluated by readers. Furthermore, various formal studies on reading comprehension (Kintsch & Keenan 1973; Graesser, Millis & Zwaan 1997; Hoover 1997) have focused on mechanical aspects such as reading rate, and did not have access to computational mechanisms for producing the texts they studied.

To study the changes in perceived text quality stemming from alterations to the underlying text generation architecture, we conducted a formal study gauging the satisfaction of subjects reading narratives. The study involved the following:

- A consistent representation mechanism which allows for the representation of characters, props, locations, actions and descriptions found in a narrative environment. Holding these entities constant for the duration of an experiment ensures that the stories seen by the study participants will have identical plots and details except for the variations cued from the experiment's parameters.

- A story generation mechanism that, when given the story representation and the experimental parameters, can produce a specified set of narratives. Our story generator, named STORYBOOK, creates narratives in the Little Red Riding Hood fairy tale domain. These narratives can be tailored to produce a variety of grammatical, lexical, and propositional effects.

- A pool of readers familiar with narratives and the writing process itself. Thus we conducted a study involving 20 upper division undergraduate students majoring in English or Communication. Each student read two distinct Little Red Riding Hood stories averaging two hours per student.

There are two primary types of comparisons upon which an evaluation of a text-producing system can focus: human text *vs.* computer text and computer text *vs.* computer text. Although there are a number of pre-existing Little Red Riding Hood texts available for comparison via the World Wide Web, formally comparing such narratives with those produced by computer presents a difficult problem: there is no known objective metric for quantitatively evaluating narrative prose in terms of how it performs *as a story*. Simple metrics exist for evaluation at the sentence level (*e.g.*, number of words, depth of embedding, *etc.*), but a narrative *per se* cannot be considered to be just a collection of sentences

that are not related to each other. In addition, because narrative is not a "deductive" domain, it cannot be evaluated in terms of *correctness* by a panel of human judges. To overcome these problems, we instead opted for a computer *vs.* computer style of evaluation that investigates whether certain architectural elements are necessary or useful when generating narrative prose.

To study the effects of different textual effects upon the readers, we implemented five versions of the STORYBOOK story generator (Callaway & Lester 2001). Because a fully interleaved experiment would have required an excessive amount of time, we required each student to compare two versions of each story rather than all five versions. Each story was identical in plot, content, and form, but differed in terms of propositions per sentence, grammatical fluency, or choice of lexical forms. The results of the study show that the participants were highly discriminative of the texts which they read, preferring some versions over others. The readers most strongly dispreferred narratives lacking important grammatical structures and greatly dispreferred those with a small number of propositions per sentence. These results have important implications for the design of literacy software.

The STORYBOOK Narrative Prose Generator

STORYBOOK is a narrative prose generator that produces narratives in the Little Red Riding Hood domain. To write stories, STORYBOOK takes a narrative plan consisting of the actors, scenes, props and temporally ordered events and descriptions as input from a narrative planner. It then evolves that narrative plan into the final text seen by the reader using a sequence of architectural components:

- *Discourse History*: When populating a story with information from a conceptual network, noun phrases must be marked for indefiniteness if they have not yet been mentioned in the story or if they are not visually available references to the character or narrator in focus. Furthermore, frequently repeating noun phrases can be pronominalized to avoid sentences like "Grandmother knew that Grandmother had asked Grandmother's daughter to send some cakes to Grandmother" rather than "Grandmother knew she had asked her daughter to send her some cakes." A discourse history tracks noun phrase concepts and allows them to be marked for definiteness or pronominalization.

- *Sentence Planner*: A sentence planner maps characters, props, locations, actions and descriptions to concrete grammatical structures in a sentential specification. Thus in the example just mentioned, "grandmother" is mapped to the main subject while "know" is mapped to the main verb, etc.

- *Revision*: Because narrative planners create their content as a series of single proposition sentences, a revision component is usually introduced to *aggregate* those small sentences (protosentences) into larger multi-proposition sentences. It is usually assumed that these larger sentences will be more readable and less choppy or visually jarring. For example, "The wolf saw her" and "She was walking down the path" might be aggregated to produce "The wolf saw her walking down the path."

- *Lexical Choice*: Narrative planners also tend to create sentences that frequently repeat the same lexical items due to efficiency concerns. To combat this, a lexical choice component performs local search to determine when one lexical item can be replaced by another. Thus instead of character dialogue where characters always introduce utterances with "said", that lexical item can be replaced by "mentioned", "whispered", "replied", etc.

- *Surface Realizer*: Once the lexical and structural content of a set of sentences has been determined, they must be converted to text. This is accomplished by checking to make sure that each sentence is grammatical, imposes linear constraints, and adds morphological changes as necessary. The result is text which can be sent to a word processor, a web browser, or saved as a text file.

The existence of these architectural modules allowed us to conduct an *architectural ablation* experiment. By selectively removing a component, the resulting text of a story will be changed in some way. The sentence planner and surface realizer are vital components; without them text cannot be produced at all. However, removing the other elements will result in text that we expect to be degraded in some fashion. Thus without the discourse history, the system will be unable to produce pronouns in a reliable way or appropriately mark nouns for definiteness. Without the revision component, the system will produce a minimal number of propositions per sentence due to the lack of clause aggregation. Finally, removing the lexical choice module will result in a decrease in the variability of the lexical forms of verbs or nouns.

Given these three architectural modules, there are 2^3 or 8 possible pairwise comparisons between the presence or absence of each component when used to produce a narrative:

1. All three components are used.
2. Only the revision module is unused.
3. Only the lexical choice module is unused.
4. Only the discourse history module is unused.
5. Only the revision module is used.
6. Only the lexical choice module is used.
7. Only the discourse history module is used.
8. None of the three components are used.

Due to the constraints on the logistics of the evaluation process, we decided to utilize only five of those pairwise comparisons: the two all-or-none approaches and the

three approaches where one specific architectural module is ablated. The remaining three unused approaches would evaluate the enhancement that each module adds to the whole rather than what is missing when each is removed. We contend this approach leads to a slightly more effective comparison, because as more modules are removed from the generation process, the resulting prose becomes progressively less desirable and thus unwanted effects from the absence of multiple architectural modules might overlap and affect a test subject's experience in ways that could not be teased apart when analyzing the data.

The ablation of these architectural modules can have a significant impact in text quality, even over very small text segments, as is shown in the following excerpts:

- Complete (Version A), with revision, lexical choice, and discourse history all turned on:

  She had not gone far when she met a wolf.

  "Hello," greeted the wolf, who was a cunning looking creature. He asked, "Where are you going?"

  "I am going to my grandmother's house," she replied.

- No Revision (Version B), with lexical choice and discourse history turned on:

  She had not gone far. She met a wolf.

  "Hello," greeted the wolf. The wolf was a cunning looking creature. He asked, "Where are you going?"

  "I am going to my grandmother's house," she replied.

- No Lexical Choice (Version C), with revision and discourse history turned on:

  She had not gone far when she met a wolf.

  "Hello," said the wolf, who was a cunning looking creature. He said, "Where are you going?"

  "I am going to my grandmother's house," she said.

- No Discourse History (Version D), with revision and lexical choice turned on:

  Little Red Riding Hood had not gone far when Little Red Riding Hood met the wolf.

  "Hello," greeted the wolf, who was the cunning looking creature. The wolf asked, "Where is Little Red Riding Hood going?"

  "Little Red Riding Hood is going to Little Red Riding Hood's grandmother's house," replied Little Red Riding Hood.

- Empty (Version E), with revision, lexical choice, and discourse history all turned off:

  Little Red Riding Hood had not gone far. Little Red Riding Hood met the wolf.

  "Hello," said the wolf. The wolf was the cunning looking creature. The wolf said, "Where is Little Red Riding Hood going?"

  "Little Red Riding Hood is going to Little Red Riding Hood's grandmother's house," said Little Red Riding Hood.

1. On an absolute scale of how good fairy tales should be in general, evaluate the story on an A–F scale (A, B, C, D, F).
2. Style: Did the author use a writing style appropriate for fairy tales?
3. Grammaticality: How would you grade the syntactic quality of the story?
4. Flow: How well did the sentences flow from one to the next?
5. Diction: How interesting or appropriate were the author's word choices?
6. Readability: How hard was it to read the prose?
7. Logicality: Did the story omit crucial information or seem out of order?
8. Detail: Did the story have the right amount of detail, or too much or too little?
9. Believability: Did the story's characters behave as you would expect?

Figure 1: Grading factors presented to readers

Evaluation Methodology

To test the StoryBook system, we created a modestly sized narrative planner (implemented as a finite state automaton containing approximately 200 states), enough to produce two stories comprising two and three pages respectively. Furthermore, we fixed the content of those stories and ran five different versions of StoryBook on each one: (A) all three components working, (B) revision turned off, (C) lexical choice turned off, (D) the discourse history turned off, and finally (E) a version with all three components turned off. This resulted in ten total narratives which we presented to our test subjects using the grading factors found in Figure 1. While versions were different in the sense that certain modules were either ablated or not, the two stories differ because they were created from two different finite state automata. Thus story #1 potentially has different characters, different events and properties, and different props than story #2 has.

A total of twenty students were selected from North Carolina State University's Departments of English and

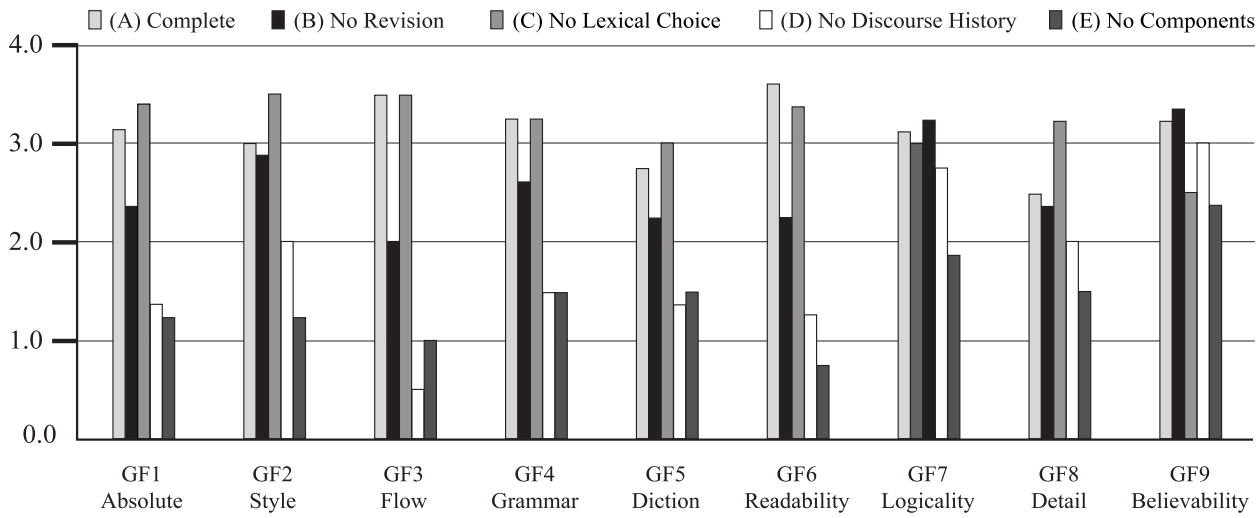


Figure 2: Means for Story #2: 4.0 scale, 8 evaluations per Version × Grading Factor × Story

Communication via first-come first-serve email notices. All of the students were registered in upper division or graduate courses in those departments. Each subject was asked to read the directions and ask for clarifications before the evaluation proceeded and was randomly assigned their evaluation task. Subjects were not informed prior to their completion of the questionnaire that the narratives were produced by computer program. Subjects were paid $25.00 for their participation.

Because each subject compared two versions of story #1 to each other and two versions of story #2 to each other, every subject saw a total of four narratives. To prevent subjects from evaluating the same types of stories in succession, we devised the following policy:

1. Each subject read four distinct story versions out of the total of five, two from each story (*e.g.*, subject #1 read versions A and B from story #1, and versions D and E from story #2). No subject read the same version twice.

2. Each version was read by the same total number of subjects (*i.e.*, each version of every story was read by 8 separate subjects).

3. Each pairwise comparison of different versions was read by two separate subjects (*e.g.*, subjects #1 and #11 both read versions A and B of story #1 and versions D and E of story #2).

4. For each pair of students reading the same two versions, the narratives were presented in opposite order (*e.g.*, subject #1 read version A first and then version B, while subject #11 read version B first followed by version A).

5. Students were randomly assigned narrative versions on a first-come first-serve basis; all students performed their evaluations within 3 hours of each other at a single location.

Subjects graded each narrative following the instructions according to an A–F scale, which we then converted to a quantified scale where A = 4.0, B = 3.0, C = 2.0, D = 1.0, and F = 0.0. The resulting scores were then tallied and averaged. The means for both stories are shown in Figure 2.

To determine the quantitative significance of the results, we performed an ANOVA test over both stories. The analysis was conducted for three independent variables (test subject, story, and version) and nine grading factors (labelled GF1 – GF9, as described in Figure 1). Because not all possible grading combinations were performed (only 80 observations, or 20 x 2 x 2, out of a possible 200, or 20 x 2 x 5, due to crossover and time constraints), we performed the mixed procedure analysis. Interactions between variables were only significant for grading factor #9 at 0.0300 for story∗version.

The results of the ANOVA analysis point to three significant classes of narratives due to the architectural design of the narrative prose generator. Table 1 indicates that the most preferred narrative class, consisting of versions A & C, were not significantly different from each other overall while they did differ significantly from all other versions (although there were similarities in particular grading factors such as GF2, *style*, between versions A & B). Interestingly, the affinity for versions A & C is strongly correlated for story #2 (Figure 2) but only weakly for story #1. A two-tailed paired t-test evaluating this difference illustrated that versions A & B were not significantly different when only story #1 was considered, but were significantly different in story #2. The opposite was true for versions A & C when the scores for each story were compared individually, even though the combined scores indicated versions A & C were not significantly different overall.

Grading Factors	GF1	GF2	GF3	GF4	GF5	GF6	GF7	GF8	GF9	ALL
COMPLETE VS. NO REV.	n.s.	n.s.	**	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
COMPLETE VS. NO L. C.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
COMPLETE VS. NO D. H.	**	*	**	**	**	**	n.s.	*	n.s.	**
COMPLETE VS. NOTHING	**	*	**	**	**	**	n.s.	n.s.	*	**
NO REV. VS. NO L. C.	*	n.s.	**	*	*	*	n.s.	n.s.	n.s.	**
NO REV. VS. NO D. H.	**	*	**	**	*	**	n.s.	n.s.	n.s.	**
NO REV. VS. NOTHING	**	n.s.	*	**	n.s.	**	n.s.	n.s.	*	**
NO L. C. VS. NO D. H.	**	**	**	**	**	**	*	**	*	**
NO L. C. VS. NOTHING	**	**	**	**	**	**	*	**	**	**
NO D. H. VS. NOTHING	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.

Table 1: Combined significance values (with Bonferroni adjustment): $*$ = p < 0.01, $**$ = p < 0.001

Discussion

Indisputably, versions D & E form the least preferred narrative class, differing quite significantly from all other versions while not differing significantly from each other. Because the architectural commonality between these two versions was the lack of a discourse history (corresponding to a lack of grammatical conformity to the expected norm, especially lack of appropriate pronominalization) while versions A, B, and C all utilized a discourse history, we conclude that this architectural component is extremely important in the design of a narrative prose generator and that any symbolic pipelined narrative prose generation system will suffer tremendous degradation in prose quality if a discourse history component is not present. In addition, we conclude that in future ablation experiments, if there is no other methodology for introducing pronominalizations, it is not even desirable to include the discourse history module as one of the components available for ablation. Effects of pronominalization and topicalization were previously studied by Hoover (1997) although that work focused on recall rates while we concentrate on expressed preferences.

As predicted in advance, the full version (Version A) scored quite well while versions lacking a discourse history (Versions D & E) scored quite poorly. A surprise in the results of the analysis was the mild preference subjects had for the version missing the lexical choice component (Version C) over the full-fledged version. While related work on word choice in spontaneous dialogues has concluded that dialogue participants tend to converge onto a limited set of words (Brennan 1996), fictional narrative by and large does not reflect the spontaneity and task-orientation reflected in such dialogues.

Upon analysis of the comments in the evaluations specifically comparing versions A & C, it became clear that one principal reason was the test subjects' belief that the increased lexical variation might prove too difficult for children to read (even though we provided no indication that the target audience was children) and thus Version A compared less favorably to Version C due to the more complex and varied words it contained. It is not clear whether a lexical choice component would play a much more significant role in subject matter where the audience was more mature.

The fact that Version B scored less favorably compared to Versions A and C indicates that revision is an important aspect of narrative prose generation. Test subjects frequently commented that Version B was "very choppy" or "didn't seem to have good grammar". These comments can be accounted for by the two main functions of the revision component: joining small sentences together and combining sentences with repetitive phrases together while deleting the repetitions. This is related to previous work in reading comprehension on propositional content. Such research (Kintsch & Keenan, 1973) has shown that reading rate increases as the number of propositions per sentence increases. Here, however, we have shown that a larger number of propositions per sentence is preferred more than a small number of propositions per sentence, although there would certainly be an upper limit.

Another important note is that there is a difference among the grading factors themselves. Grading factors 2-7 (style, flow, grammar, diction, readability and logicality) directly relate to elements governed by the parameters and rules of the various architectural components of the narrative prose generator. However, grading factors #8 and #9 (detail and believability) are more closely related to the content of the plot line, and as such could be expected to remain relatively constant since the content of the narratives was held constant across all versions of each story. Given that the perceptions of the test subjects might have "carried over" from their responses to previous questions, a future evaluation might randomize the order in which these questions are asked to see if this effect persists.

Finally, there appears to be a link between the appeal of the story content itself and the increase in the absolute (GF #1) and total means for versions A, B, and C. Story #1 is a "classic" Brothers' Grimm fairy tale in the sense that it typically has a gruesome ending that serves as a behavioral warning to young children. Thus our story #1 ends with the wolf devouring Little Red Riding Hood

and her grandmother. More modern stories have happier endings, however, and this is reflected in our story #2 which ends with a woodcutter killing the wolf and extracting the unharmed Little Red Riding Hood and her grandmother from the wolf's stomach. A large number of our test subjects, worried about the potential impact on children, complained about the "horrible" ending of story #1 in their written comments and this reader bias appears to have affected the overall grading scores.

Future Work

The existence of a computational system for generating complete narratives while providing access to the fundamental linguistic structure offers superb opportunities for future experimentation. Very fine-grained manipulation of texts becomes possible on a large scale; for example, within the discourse history, it is possible to run ablation experiments involving subject pronouns vs. object pronouns, correct vs. incorrect reflexive pronouns, random vs. ambient definite noun phrase marking, among many others.

Acknowledgements

The authors wish to thank Joy Smith of NC State University for her help with the statistical analysis and the anonymous reviewers for their extremely helpful comments. Support for this work was provided by the IntelliMedia Initiative of North Carolina State University.

References

Allen, J., Miller, B., Ringger, E., & Sikorski, T. (1996). Robust understanding in a dialogue system. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, (pp. 62–70). Santa Cruz, CA.

Brennan, S. (1996). Lexical entrainment in spontaneous dialog. In *Proceedings of the International Symposium on Spoken Dialogue*. Philadelphia, PA.

Callaway, C., & Lester, J. (1997). Dynamically improving explanations: A revision-based approach to explanation generation. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, (pp. 952–958). Nagoya, Japan.

Callaway, C., & Lester, J. (2001). Narrative prose generation. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, in press. Seattle, WA.

Graesser, A. C., Millis, K. K. & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology*, **48**: 163–189.

Hoover, M. L. (1997). Effects of textual and cohesive structure on discourse processing. *Discourse Processes*, **23**: 193–220.

Kintsch, W. & Keenan, J. M. (1973). Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, **5**:257–274.

Lang, R. R. (1997). *A formal model for simple narratives*. Doctoral Dissertation, Department of Computer Science, Tulane University. New Orleans, LA.

Lebowitz, M. (1985). Story-telling as planning and learning. *Poetics*, **14**(3): 483–502.

Lester, J. & Porter, B. (1997). Developing and empirically evaluating robust explanation generators: The KNIGHT experiments. *Computational Linguistics*, **23**(1): 65–101.

Meehan, J. (1977). Tale-Spin, an interactive program that writes stories. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*. Cambridge, MA.

Robin, J. & McKeown, K. (1995). Empirically designing and evaluating a new revision-based model for summary generation. *Artificial Intelligence*, **85**(1–2).

Smith, R. & Hipp, D. R. (1994). *Spoken natural language dialog systems*. Cambridge, Massachusetts: Oxford University Press.

Young, R. M. (1999). Using Grice's Maxim of Quantity to select the content of plan descriptions. *Artificial Intelligence*, **115**: 215–256.

How Nouns and Verbs Differentially Affect the Behavior of Artificial Organisms

Angelo Cangelosi (acangelosi@plymouth.ac.uk)
PION Plymouth Institute of Neuroscience and Centre for Neural and Adaptive Systems
University of Plymouth
Drake Circus, Plymouth, PL4 8AA, UK

Domenico Parisi (parisi@ip.rm.cnr.it)
Institute of Psychology, National Research Council
Viale Marx 15, Rome, 00137, Italy

Abstract

This paper presents an Artificial Life and Neural Network (ALNN) model for the evolution of syntax. The simulation methodology provides a unifying approach for the study of the evolution of language and its interaction with other behavioral and neural factors. The model uses an object manipulation task to simulate the evolution of language based on a simple verb-noun rule. The analyses of results focus on the interaction between language and other non-linguistic abilities, and on the neural control of linguistic abilities. The model shows that the beneficial effects of language on non-linguistic behavior are explained by the emergence of distinct internal representation patterns for the processing of verbs and nouns.

Modeling the Evolution of Language

The recent development of computational evolutionary models (Wiles & Hallinan, in press) has contributed to the rebirth of interest in the origin and evolution of language. Computational models can directly simulate the evolution of communication and the emergence of language in populations of interacting organisms (Cangelosi & Parisi, in press; Dessalles & Ghadakpour, 2000; Steels, 1997). Various simulation approaches are used such as communication between rule-based agents (Kirby, 1999), recurrent neural networks (Batali, 1994; Ellefson & Christiansen, 2000), robotics (Kaplan, 2000; Steels & Vogt, 1997), and internet agents (Steels & Kaplan, 1999).

Artificial Life Neural Networks (ALNN) are neural networks controlling the behavior of organisms that live in an environment and are members of evolving populations of organisms. ALNN models have been used to simulate the evolution of language (Cangelosi & Parisi, 1998; Cangelosi, 1999; Cangelosi & Harnad, in press; Parisi, 1997). For example, in Cangelosi and Parisi's (1998) model organisms evolve a shared lexicon for naming different types of foods. Communication signals are processed by neural networks with genetically inherited connection weights and the signals evolve at the population level using a genetic algorithm with no changes during an individual's lifetime.

ALNN models provide a unifying methodological and theoretical framework for cognitive modeling because of the use of both evolutionary and connectionist techniques and the interaction of the organisms with a simulated ecology (Parisi, 1997). All behavioral abilities (e.g., sensorimotor skills, perception, categorization, language) are controlled by the same neural network. This unified framework permits the study of various factors affecting language evolution, such as the differences between genetic and learned communication systems, the adaptive role of both simple and compositional languages, the neural control of language, the reciprocal influences between language and cognition.

Emergence of compositional languages: verbs and nouns

The evolutionary emergence of messages that combine two linguistic signals has been studied with ALNN models. In Cangelosi and Parisi's (1998) model, organisms communicate using simple signals that are genetically inherited. In an extension of the model, word combination and language learning were introduced to simulate the emergence of compositional languages (Cangelosi, 1999; in press). The organisms' neural networks had two linguistic winner-takes-all output clusters so that two "words" were simultaneously uttered to name foods (different types of mushrooms). Parents acted as linguistic teachers of their offspring. Children learned to name foods by imitating their parents' descriptions using an error backpropagation algorithm.

The simulation results showed that about 60% of the populations evolved optimal languages, i.e., languages in which each category of food was correctly identified and named. In the remaining populations, only one category out of six was misclassified. Evolved languages were classified into three types: (1) Single-

word, where the units in only one cluster are enough to differentiate all mushrooms; (2) Word-combination, where symbols from both clusters are needed to discriminate mushrooms; (3) Verb-Noun, where the units in one cluster are systematically associated with high-order categories (e.g., "verbs" for approaching/avoiding) and the other cluster is used for differentiating sub-categories (e.g., "nouns" for mushrooms of different color).

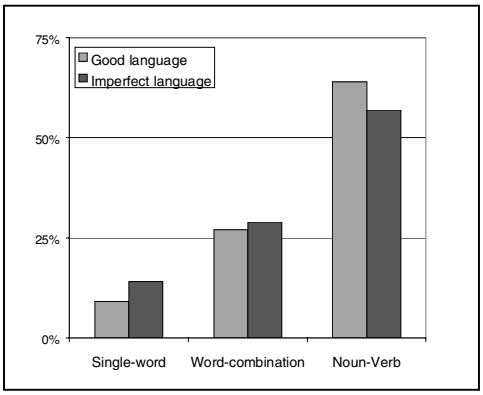


Figure 1: Distribution of languages in the 18 simulations with communication (at generation 400).

The distribution of language types (Figure 1) showed that there is a strong evolutionary tendency to evolve compositional languages, where the syntactic structure of messages reflects the hierarchical classification of mushrooms. In fact, the most frequent (e.g., 64% of good languages) combinatorial structure is that of predicate-argument, resembling a "verb-noun" sentence.

Behavioral and Neural Factors in the Evolution and Acquisition of Language and Syntax

We will now focus on some issues about the acquisition and use of language, and on their relations with language evolution studies. These issues regard the interaction between language and other behavioral abilities, the stages of the acquisition and evolution of syntax, and the organization of neural representations in language processing. The first issue is quite an important and old one: How does language affect, and how is it affected by, other cognitive and behavioral abilities? Various language origin theories stress the importance of pre-existing sensorimotor knowledge for effective evolution of linguistic skills. For example, Rizzolatti and Arbib (1998) proposed a motor theory of language evolution based on imitation skills. Steels (2000) showed how his robotics models of language evolution support this theory. In Cangelosi and Parisi's (1998) ALNN model, they showed how language evolution relies on the evolution of basic cognitive abilities such as categorization. The dependence of language on previous sensorimotor skills, and the effects of language on this behavior will be looked at in the models presented here.

Researchers interested in both the evolution and the acquisition of language, are primarily concerned with the early stages of the development of linguistic abilities. In particular they focus on the transition from a non-linguistic stage where sensorimotor abilities dominate to a phase in which language and other high order cognitive skills emerge and take control of cognitive development. Although little empirical evidence is available for language evolution, data on language acquisition strongly support the conclusion that children learn nouns before verbs (Brooks & Tomasello, 1999). They handle nouns at around 18 months, while verbs are acquired later, from around 24 months. Verbs seem to follow a more gradual acquisition pattern, passing through an intermediate stage called "verb islands" (Tomasello, 1992). We will use data from our simulations to look for similar learning patterns in language evolution.

The investigation of the neural control of nouns vs verbs has been the focus of some interesting neuropsychological and brain imaging studies. For example, Caramazza and Hillis (1991) looked at the brain representation of noun and verbs in patients with brain lesions. Martin, Haxby, Lalonde, Wiggs & Ungerleider (1995) used PET to show that cortical sensory areas are active when the color word of an object is retrieved, while motor areas are involved in the processing of action words. ALNNs permit the investigation of internal representations involved in the processing of different syntactic classes such as nouns and verbs.

In the next section we will describe a new ALNN model of the evolution of syntax, specifically the verb-noun syntactic rule. This simulation will be use to study in detail the interaction between linguistic abilities and other behavioral and neural factors.

Evolution of Verb-Noun Languages

The ALNN model described in Cangelosi, 1999 (cf. also Cangelosi, in press) showed a significant tendency to evolve compositional languages made up of verb-noun messages. To study the differences between verbs and nouns and how verb-noun languages affect and are affected by other behavioral, cognitive, and neural factors, a new model with a pre-defined compositional language will be used. The language includes four simple linguistic signals (words), two nouns and two verbs. Nouns are defined as linguistic signals that covary with the visual input. Verbs are defined as linguistic signals that covary with the action of the organism. Messages can include only a noun or only a verb or they can be a combination of a noun and a verb.

The Model

The task used in the simulation is an object manipulation task (Schlesinger & Barto, 1999). At any given time the organism is grasping an object with its hand and it either pulls the object toward itself or it pushes the object away from itself. Two different objects are used, a vertical bar (object A) and a horizontal bar (object B). The object is perceived through a retina of 5x5=25 cells corresponding to 25 visual input units. The object occupies either three vertical cells or three horizontal cells in one of 9 possible locations in the retina. Hence, an object is encoded as a pattern of 25 bits with three 1s and twenty-two 0s. In addition to this visual input from the retina the organism's neural network receives a proprioceptive input encoding the current position of the organism's two-segment arm. This input is encoded in 4 input units, with units encoding proprioceptive information about the two pairs of muscles (extensor and contractor) of each of the two arm segments.

In the simulations with language the neural network includes 4 more input units encoding linguistic signals. Four linguistic signals are used, two nouns and two verbs, and they are localistically encoded in the 4 linguistic input units One noun designates the vertical object and a different noun designates the horizontal object. One verb designates the action of pushing and the other verb the action of pulling the object. In different occasions the organism can perceive only a noun or only a verb or both a noun and a verb.
There are two layers of hidden units that receive information from the input units and pass it to the 4 output units (Figure 2). The output units control the extension/contraction of the four arm muscles.

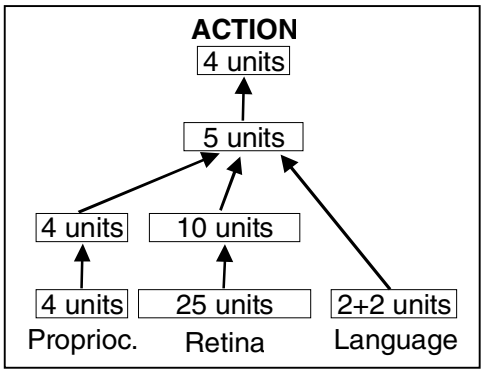


Figure 2 – The organism's neural network for the object manipulation task

The connection weights allowing the neural network to appropriately manipulate the two objects are developed using a genetic algorithm. At the beginning of a simulation 80 genotypes are randomly generated each encoding the connection weights of a single individual. These 80 individuals constitute the first generation. The 20 best individuals are selected for reproduction, with each individual generating 4 offspring with the same genotype (connection weights) of its single parent except for the addition of some random changes to some of the weights (random mutations). The process is repeated for 2000 generations.

Three experimental conditions were used. In the first condition, called "No-Language", an organism lives for a single epoch consisting of a total of 360 input/output mappings or moves (2 object types x 9 positions x 20 moves per task). Only the retina and the proprioceptive information are provided as input to the network. When the organism sees object A, it always has to push it away from itself; when it sees object B, it has to pull it towards itself. The fitness formula computes the total number of tasks successfully completed.

The second experimental condition is called "Late-Language". At generation 1000 a copy of the populations of the No-Language condition is made. From this generation onwards the organisms have a longer lifetime and they are exposed to language. Ten new epochs with language are added to an individual's lifetime, which therefore now includes 11 epochs, 10 with language and 1 without language. In 5 of the linguistic epochs an individual receives both the linguistic input and the retina and proprioceptive inputs, whereas in the remaining 5 epochs only the linguistic input and the proprioceptive input are present and the retina input is shut off. The 5 linguistic epochs are as follows: (1) add noun of the object, (2) add verb corresponding to the default action (push object A or pull object B), (3) add verb for opposite action (pull object A or push object B), (4) add both noun and default verb, and (5) add both noun and opposite verb. The various epochs are experienced by an organism in a random sequence. The same fitness formula is used as in the No-language case except that in the epochs when the opposite verb is used, the organism's action must reflect what the verb says, not what the object type would suggest by default.

In the third experimental condition, "Early-Language", organisms are exposed to all 11 epochs from the beginning of the simulation, i.e., from the first generation. For each condition, 20 replications of the simulations were run.

Results and Discussion

The average performance of the organism in the three simulations is reported in Figure 3. For the two linguistic conditions, only the curve of the performance in the epoch with no linguistic input is reported, to allow a direct comparison among the three conditions. The No-language fitness curve grows until it stabilizes at around 15.8 successfully completed epochs. In the

Late-Language condition, at generation 1001 the population goes through a significant drop in performance. This appears to be due to the fact that the linguistic input reaches the hidden units through random weights that disturb the previous good performance. However, the behavior gradually improves and from around generation 1400 Late-Language organisms outperform No-Language organisms. The final number of successful tasks is 16.6 for the Late-Language condition. In contrast with this, the performance of the Early-Language population is less good than that of both the No-Language and the Late-Language populations (14.4).

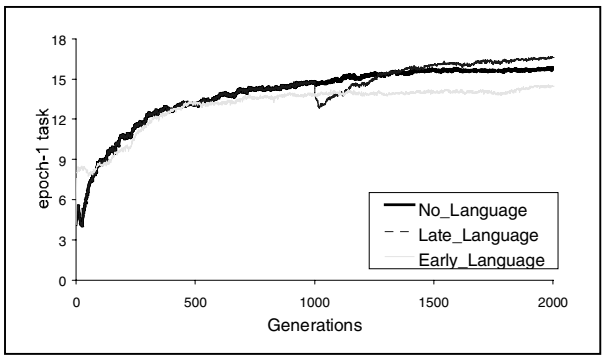


Figure 3 – Performance in epoch 1 (task without linguistic input) in the three experimental conditions

These results suggest an interesting hypothesis on language evolution and the interaction between linguistic and cognitive skills. To be adaptive language must be introduced at a later stage, after the cognitive abilities upon which it will be grounded have fully evolved. In this condition language has a beneficial influence on nonlinguistic behavior. If the evolutionary scenario involves both the practical task of pushing or pulling objects and the processing of linguistic signals from the beginning, it is more difficult to evolve populations with optimal performance in the practical task. Notice that if language is introduced later so that it can exploit the already existing (nonlinguistic) cognitive skills, the beneficial effects of language on nonlinguistic performance are observed not only when language is used together with nonlinguistic input (the language epochs) but also when there is no language and the organism is responding only to nonlinguistic input.

We will now focus on the Late-Language simulation to better understand why language has beneficial effects on nonlinguistic behavior and to analyze the differences between the two different classes of linguistic signals: nouns and verbs.

The 11 epochs of the Late-Language simulation can be grouped into 4 categories: (1) No-language, (2) Noun-only (the 2 epochs with and without retina input), (3) Verb-only (the four epochs with/without retina and with default/opposite verbs), and (4) Verb+Noun (the four epochs with/without retina and with default/opposite verbs).

Figure 4 shows the average performance for the three linguistic categories (categories 2-4) from generation 1000 to generation 1300. In the early generations, right after language has been introduced (from generation 1000 to generation 1100) the organisms' performance in the Noun-only epochs is higher than that of Verb-only and of Noun+Verb. Organisms learn to use nouns earlier than verbs to benefit their nonlinguistic performance. However, 100 generations later the disadvantage of the verb epochs disappears. Indeed, the performance for Verb-only and Verb+Noun epochs becomes stably better than that of Noun-only epochs.

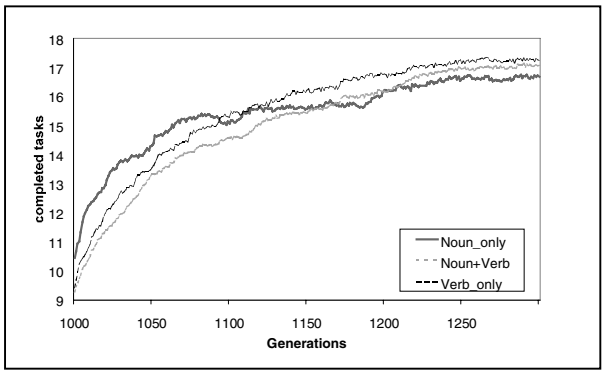


Figure 4 – Evolution of noun and verb use in the Late-Language simulation

The earlier advantage of nouns vs verbs can be explained by the fact that in the Noun-only epochs the task is consistent with what has been already learned without language up to generation 1000. Given this consistency with prelinguistic experience, nouns are easier to learn and they can benefit nonlinguistic performance earlier than verbs. On the contrary, with verbs organisms must learn to ignore some of the previously learned knowledge. When an opposite verb asks the organism to produce a new behavior (e.g., pull object A instead of pushing it, as previously learned) this is initially difficult to learn. Therefore, verbs can acquire an adaptive advantage only in later stages of language acquisition, when noun use has reached a good level of performance and stabilization and the individual can understand the more flexible nature of verbs, which can typically be predicated of a variety of arguments. This hypothesis could also explain the different stages of acquisition of nouns and verbs in children (Tomasello & Brooks, 1999). Verbs need a stable class of nouns to fully develop the potential and flexibility of their predicate-argument structure.

The Late-Language simulation can also be used to look at some aspects of the neural processing of language. To this purpose we analyzed the activation patterns in the second layer of hidden units (Figure 2), where sensory (retina+proprioception) and linguistic information come together and they both can have a role in determining the organism's motor behavior encoded in the output units. We used the activation patterns observed in these hidden units in the first cycle of each of the 18 motor tasks (9 for object A and 9 for object B). Each activation pattern can be represented as a point in the activation hyperspace of the hidden layer, with the 9 points corresponding to object A making up a "cloud" of points and the 9 points of object B making up another "cloud". We measured both the Euclidean distance between the centers of the two clouds and the size of each cloud as the average distance of the 9 points from the center of the cloud. (The points corresponding to objects/positions incorrectly handled were excluded from these calculations. On average, only 0.25 objects per epoch were misclassified.) The idea is that the successful accomplishment of the task requires that the different input patterns corresponding to the same object in different positions be internally represented as similarly as possible (small clouds) while the input patterns corresponding to the two different objects be represented as differently as possible (great distance between the two clouds).

The between-cloud distances and the sizes of the two clouds were computed for all 11 epochs. Then the data were averaged over the 4 categories of epochs: No-Language, Noun-only, Verb-only, and Noun+Verb. Figure 5 reports the average withn- and between-cloud distances at generation 2000. The between-cloud distances show a progressive increase from the No-language to the linguistic conditions. In an ANOVA test, these differences are statistically significant, except between the pair Verb-Only and Noun+Verb. A similar, but inverted, pattern of results is found for cloud size. The average size of a cloud decreases from the No-language to the linguistic conditions.

That language optimizes the representation of categories (i.e. increasing between-category distances and decreasing within-category sizes) has already been shown in other models (Cangelosi & Harnad, in press). What this model shows for the first time is that there are significant differences also between the three linguistic conditions, in particular between nouns and verbs. When the network is processing verbs, the size and distance of clouds is even better than when it is processing nouns.

How can we explain that verbs have even greater beneficial effects on nonverbal behavior than nouns? As we have shown, the beneficial effect of linguistic signals on nonlinguistic performance is due to the fact that linguistic signals induce better internal representations of reality. In our model, reality is internally represented in the neural network as the activation patterns observed in the higher layer of hidden units. The addition of language increases the distance between the two clouds of points (activation patterns) representing the two objects and decreases the size of the two clouds of points each representing one object. The language-modified clouds make it easier for the organism to select the appropriate action in response to the input. However, what is critical in internally representing reality is not to faithfully reflect the properties of the input but rather to prepare the motor output with which the organism must respond to the input. If the organism must be able to respond to the same object in different occasions with two different actions (push or pull) verbs are better than nouns in shaping the internal representations because while nouns covary with objects verbs covary with actions.

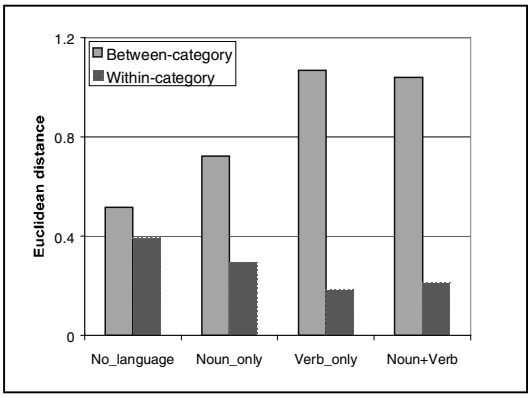


Figure 5 – Inter- and intra-categorical distances for the hidden representations at generation 2000.

Conclusion

The present model focuses on the evolution of an innate language understanding ability for a language made up of nouns and verbs. Notwithstanding its obvious limitations, the models sheds some light on the reciprocal influences between language and nonlinguistic cognition, on the differences between nouns and verbs, and on the internal organization of neural networks that use language in an ecological context. Language has a beneficial effect on nonlinguistic cognition if it emerges on already existing basis of nonlinguistic skills, but not if it evolves together with them. The basis for this beneficial influence of language on behavior appears to be that language produces better internal representations of reality. That is, more similar representations of different situations that must be responded to with the same action, and more different internal representations of similar situations that must be responded to with different behaviors. Furthermore, verbs have a more beneficial effect on behavior than nouns because verbs,

by their nature, tend to covary with the organism's actions while nouns tend to covary with the objects of reality that may be responded to with different actions in different occasions.

In this paper we have also done some comparisons between the computational model of language evolution and the literature on children's language acquisition and on neural processing of verbs and nouns. We are currently working on an extension of the object manipulation model to understand better the relations between language processing and sensorimotor knowledge (Martin et al, 1995). All in all, we believe this is a fruitful approach to the investigation of various adaptive, behavioral, and neural factors involved in the origin and evolution of language.

Acknowledgments

Angelo Cangelosi's work for this paper was partially funded by an UK EPSRC Grant (GR/N01118).

References

Batali, J. (1994). Innate biases and critical periods: Combining evolution and learning in the acquisition of syntax. In R. Brooks & P. Maes (Eds), *Artificial Life IV* (pp. 160-171), Cambridge, MA: MIT Press.

Cangelosi, A. (in press). Evolution of communication and language using signals, symbols, and words. *IEEE Transactions on Evolutionary Computation*.

Cangelosi, A. (1999). Modeling the evolution of communication: From stimulus associations to grounded symbolic associations. In D. Floreano et al. (Eds.), *Proceedings of ECAL99 European Conference on Artificial Life* (pp. 654-663), Berlin: Springer-Verlag.

Cangelosi, A., & Harnad, S. (in press). The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication*.

Cangelosi, A. & Parisi, D. (in press). *Simulating the Evolution of Language*. London: Springer-Verlag.

Cangelosi, A, & Parisi, D. (1998). The emergence of a language in an evolving population of neural networks. *Connection Science*, 10, 83-97.

Caramazza, A., & Hillis, A.E. (1991). Lexical organization of nouns and verbs in the brain. *Nature*, 349, 788-900.

Deacon, T.W. (1997). *The Symbolic Species: The Coevolution of Language and Human Brain*, London: Penguin.

Dessalles, J., & Ghadakpour, L. (Eds.) (2000). *Proceedings of the 3rd International Conference on the Evolution of Language*. Paris: ENST Press.

Di Ferdinando, A., Calabretta, R., & Parisi, D. (2001). Evolving modular architectures for neural networks. In R. French & J. Sougné (Eds.), *Proceedings of the Sixth Neural Computation and Psychology Workshop Evolution, Learning, and Development*, London: Springer Verlag.

Ellefson, M.R. & Christiansen, M.H. (2000). Subjacency constraints without universal grammar: Evidence from artificial language learning and connectionist modeling. In *The Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 645-650). Mahwah, NJ: Lawrence Erlbaum.

Kaplan, F. (2000). Talking AIBO: First experimentation of verbal interactions with an autonomous four-legged robot. In A. Nijholt, D. Heylen, & K. Jokinen (Eds.). *Learning to Behave: Interacting agents. CELE-TWENTE Workshop on Language Technology* (57-63).

Kirby, S. (1999). Syntax out of learning: The cultural evolution of structured communication in a population of induction algorithms. In D. Floreano et al. (Eds.), *Proceedings of ECAL99 European Conference on Artificial Life* (pp. 694-703), Berlin: Springer-Verlag.

Martin, A., Haxby, J.V., Lalonde, F.M., Wiggs, C.L., & Ungerleider, L.G. (1995). Discrete cortical regions associated with knowledge of color and knowledge of action. *Science*, 270, 102-105.

Parisi, D. (1997). An Artificial Life approach to language. *Mind and Language*, 59, 121-146.

Rizzolatti, G. & Arbib, M. (1998). Language within our grasp. *Trends in Neuroscience*, 21, 188-194.

Schlesinger, M., & Barto, A. (1999). Optimal control methods for simulating the perception of causality in young infants. In M. Hahn & S.C. Stoness (Eds.), *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society* (pp. 625-630). New Jersey: Lawrence Erlbaum.

Steels, L. (1997) The synthetic modeling of language origins. *Evolution of Communication*, 1, 1-37.

Steels, L. (2000) Mirror neurons and the action theory of language origins. *Proceedings of Architectures of the Mind, Architectures of the Brain*.

Steels, L., & Kaplan, F., (1999). Collective learning and semiotic dynamics. In D. Floreano et al. (Eds.), *Proceedings of ECAL99 European Conference on Artificial Life* (pp. 679-688), Berlin: Springer-Verlag.

Steels, L., & Vogt, P. (1997). Grounding adaptive language games in robotic agents. In P. Husband & I. Harvey (Eds). *Proceedings of the Fourth European Conference on Artificial Life* (pp. 474-482), London: MIT Press.

Tomasello, M., & Brook, P.J. (1999). Early syntactic development: A Construction Grammar approach. In M. Barrett (Ed.), *The Development of Language* (161-190). Philadelphia, PA: Psychology Press

Wiles, L., & Hallinan, J.S. (in press). Evolutionary computation and cognitive science: Modeling evolution and evolving models. *IEEE Transactions on Evolutionary Computation*, Special Issue on Evolutionary Computation and Cognitive Science.

Learning Grammatical Constructions

Nancy C. Chang (nchang@icsi.berkeley.edu)
International Computer Science Institute
1947 Center Street, Suite 600, Berkeley, CA 94704 USA

Tiago V. Maia (tmaia@cse.buffalo.edu)
State University of New York at Buffalo
226 Bell Hall, Buffalo, NY 14260-2000 USA

Abstract

We describe a computational model of the acquisition of early grammatical constructions that exploits two essential features of the human grammar learner: significant prior conceptual and lexical knowledge, and sensitivity to the statistical properties of the input data. Such principles are shown to be useful and necessary for learning the structured mappings between form and meaning needed to represent phrasal and clausal constructions. We describe an algorithm based on Bayesian model merging that can induce a set of grammatical constructions based on simpler previously learned constructions (in the base case, lexical constructions) in combination with new utterance-situation pairs. The resulting model shows how cognitive and computational considerations can intersect to produce a course of learning consistent with data from studies of child language acquisition.

Introduction

This paper describes a model of grammar learning in which linguistic representations are grounded both in the conceptual world of the learner and in the statistical properties of the input. Precocity on both fronts has previously been exploited in models of lexical acquisition; we focus here on the shift from single words to word combinations and investigate the extent to which larger phrasal and clausal constructions can be learned using principles similar to those employed in word learning. Our model makes strong assumptions about prior knowledge – both ontological and linguistic – on the part of the learner, taking as both inspiration and constraint the course of development observed in crosslinguistic studies of child language acquisition.

After describing our assumptions, we address the representational complexities associated with larger grammatical constructions. In the framework of Construction Grammar (Goldberg, 1995), these constructions can, like single-word constructions, be viewed as mappings between the two domains of *form* and *meaning*, where form typically refers to the speech or text stream and meaning refers to a rich conceptual ontology. In particular, they also involve relations among multiple entities in both form (e.g., multiple words and/or phonological units) and meaning (multiple participants in a scene), as well as mappings across relations in these two domains. We introduce a simple formalism capable of representing such relational constraints.

The remainder of the paper casts the learning problem in terms of two interacting processes, construction hypothesis and construction reorganization, and presents an algorithm based on Bayesian model merging (Stolcke, 1994) that attempts to induce the set of constructions that best fits previously seen data and generalizes to new data. We conclude by discussing some of the broader implications of the model for language learning and use.

Conceptual and lexical prerequisites

Children learning their earliest word combinations bring considerable prior knowledge to the task. Our model of grammar learning makes several assumptions intended to capture this knowledge, falling into two broad categories: representational requirements for ontological knowledge; and the ability to acquire lexical mappings.

Infants inhabit a dynamic world of continuous percepts, and how they process and represent these fluid sensations remains poorly understood. By the time they are learning grammar, however, they have amassed a substantial repertoire of concepts corresponding to people, objects, settings and actions (Bloom, 1973; Bloom, 2000). They are also competent event participants who have acquired richly structured knowledge about how different entites can interact (Tomasello, 1992; Slobin, 1985), as well as sophisticated pragmatic skills that allow them to determine referential intent (Bloom, 2000).

Few computational models of word learning have addressed the general problem of how such sensorimotor and social-cultural savvy is acquired. Several models, however, have tackled the simpler problem of how labels (either speech or text) become statistically associated with concepts in restricted semantic domains, such as spatial relations (Regier, 1996), objects and attributes (Roy and Pentland, 1998), and actions (Bailey et al., 1997; Siskind, 2000). Such models assume either explicitly or implicitly that lexical items can be represented as *maps* (i.e., bidirectional associations) between representations of *form* and *meaning* that are acquired on the basis of input associations.[1] Most of these also produce word senses whose meanings exhibit category and similarity

[1]Typically, supervised or unsupervised training is used to induce word categories from sensorimotor input, which is described using continuous or discrete features; models vary in the degree of inductive bias present in the input feature space.

effects like those known to be pervasive in human cognition (Lakoff, 1987): concepts cluster into categories with prototype structure and graded category membership.

For our current spotlight on the acquisition of grammatical structures, we will make a similar set of simplifying assumptions. We do not attempt to model the complex reasoning and inference processes needed to infer the appropriate intended meaning of an utterance in context; rather, we take as input a representation of the inferred meaning in a given situational context. We also assume that lexical maps like those produced by the word-learning models described above are available as input to the grammar-learning process.

For present purposes, the precise variant of word learning is not at issue, as long as several representational requirements are met. Lexical maps should facilitate the identification of similar concepts and provide some basis for generalization. They must also be able to capture the kinds of event-based knowledge mentioned above: the meanings of many early words and constructions involve multiple entities interacting within the context of some unified event (Bloom, 1973) or basic scene (Slobin, 1985). Fortunately, these representational demands have long been recognized in the context of adult constructions, and semantic descriptions based on *frames* relating various participant *roles* have been developed by, e.g., the Berkeley FrameNet project (Baker et al., 1998). Frame-based representations can capture the relational structure of many concepts, including not only early sensorimotor knowledge but also aspects of the surrounding social and cultural context.

It will be convenient to represent frames in terms of individual role bindings: Throw.thrower:*Human* and Throw.throwee:*Object* together bind a Throw frame with a *Human* thrower acting on an *Object* throwee. Note that although this representation highlights relational structure and obscures lower-level features of the underlying concepts, both aspects of conceptual knowledge will be crucial to our approach to language learning.

In the current model, ontological knowledge is represented with an inheritance hierarchy in which frames are represented as feature structures (i.e., attribute-value matrices) and role bindings are handled by unification. Our initial set of constructions contains a number of lexical form-meaning maps, where for simplicity we further constrain these to be mappings from orthographic forms to feature-structure meanings, as in Bailey (1997).

We now turn to the representationally more complex case of grammatical constructions, before addressing how such constructions are learned.

Grammatical Constructions

We base our representations of grammatical knowledge on ideas from Construction Grammar (Goldberg, 1995) and Cognitive Grammar (Langacker, 1987). In these approaches, larger phrasal and clausal units are, like lexical constructions, pairings of form and meaning. A key observation in the Construction Grammar tradition is that the meaning of a sentence may not be strictly predictable from the meaning of its parts; the syntactic pattern itself may also contribute a particular conceptual framing. For example, the CAUSED-MOTION construction underlying *Pat sneezed the napkin off the table* imposes a causative reading on the typically non-causative verb *sneeze*, and the need for an agentive recipient in the DITRANSITIVE construction renders *Harry kicked the door the ball* somewhat anomalous.

On this account, syntactic patterns are inextricably linked with meaning, and grammaticality judgments are rightly influenced by semantic and pragmatic factors. The interpretation and acceptability of an utterance thus depends not only on well-formedness conditions but also on the structure of the language user's conceptual ontology and on the situational and discourse context.

The main representational complexity introduced with these multiword constructions is the possibility of structure in the form pole. As mentioned above, although individual lexical items can evoke complex frames with multiple participant roles (e.g., *bye-bye*, *baseball*), the actual mapping between the form and meaning pole is necessarily straightforward. With multiple form units available, however, additional structures arise, both within the form pole itself and, more significantly, in the *relational correlations* between the form and meaning poles.[2] That is, a multiword construction may involve a more complex, *structured map* between its form and meaning poles, with maps between form and meaning relations whose arguments are also mapped.

In addition to the sound patterns of individual words, the form pole includes intonational contours, morphological inflections and word order. As with single words, the meaning pole encompasses the much larger set of frame-based conceptual knowledge. The constructional mapping between the two domains typically consists of a set of form relations (such as word order) corresponding to a set of meaning relations (such as role-filler bindings).

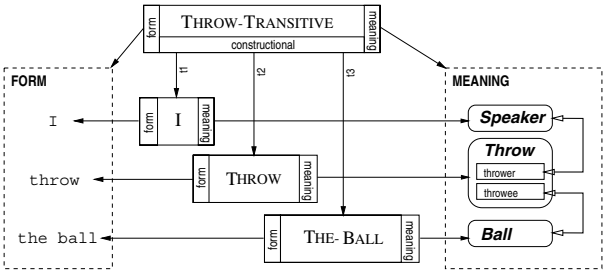


Figure 1: A constructional analysis of the sentence, *I throw the ball*, with form elements at left, meaning elements at right and some constituent constructions linking the two domains in the center.

As an example, Figure 1 gives an iconic representation of some of the possible constructions involved in an analy-

[2] See Gasser and Colunga (2000) for arguments that the ability to represent relational correlations underlies infants' reputed aptitude for statistically driven learning of concrete and abstract patterns.

sis of *I throw the ball*. The lexical constructions for I, THROW and THE-BALL[3] all have simple poles of both form and meaning. But besides the individual words and concepts involved in the utterance, we have several word order relationships (not explicitly shown in the diagram) that can be detected in the form domain, and bindings between the roles associated with Throw and other semantic entities (as denoted by the double-headed arrows within the meaning domain). Finally, the larger clausal construction (in this case, a verb-specific one) has constituent constructions, each of which is filled by a different lexical construction.[4] Crucially, the clausal construction serves to associate the specified form relations with the specified meaning relations, where the arguments of these relations are already linked by existing (lexical) maps. For example, the fact that the I construction's form pole comes *before* the THROW construction's form pole means that the meaning pole of I (i.e., the speaker in the situation) fills the thrower role in the Throw frame.

A more formal representation of the THROW-TRANSITIVE construction is given in Figure 2. For current purposes, it is sufficient to note that this representation captures the constituent constructions, as well as constraints on its formal, semantic and constructional elements. Each constituent has an alias used locally to refer to it, and subscripts f and m are used to denote the constituent's form and meaning poles, respectively. A designation constraint (in Langacker's (1987) sense) specifies a meaning type for the overall construction.

construction THROW-TRANSITIVE
  constituents:
    construct t1 **of meaning type** *Human*
    construct t2 **of type** THROW
    construct t3 **of meaning type** *Object*
  formal constraints:
    $t1_f$ *before* $t2_f$
    $t2_f$ *before* $t3_f$
  semantic constraints:
    $t2_m$.thrower $\longleftrightarrow$ $t1_m$
    $t2_m$.throwee $\longleftrightarrow$ $t3_m$
designates $t2_m$

Figure 2: Formal representation of the THROW-TRANSITIVE construction, with separate blocks listing constituent constructions, formal constraints (e.g., word order) and semantic constraints (role bindings).

Although this brief discussion necessarily fails to do justice to Construction Grammar and related work, we hope that it nevertheless conveys the essential representational demands on the structures to be learned.

[3]The definite determiner *the* explicitly depends on a representation of the situational and discourse context that supports reference resolution. For simplicity, we will ignore the internal structure of "the ball" and treat it as an unstructured unit.

[4]This example, like the rest of those in the paper, is based on utterances from the CHILDES corpus (MacWhinney, 1991) of child-language interaction.

Learning Constructions

We can now specify our construction learning task: Given an initial set of constructions C and a sequence of new training examples, find the best set of constructions C' to fit the seen data and generalize to new data. In accord with our discussion of conceptual prerequisites, a training example is taken to consist of an utterance paired with a representation of a situation, where the former is a sequence of familiar and novel forms, and the latter a set of frame-based conceptual entities and role bindings representing the corresponding scene.

Previous work on Bayesian model merging (Stolcke, 1994; Bailey et al., 1997) provides a suitable starting point. In that framework, training data is first incorporated, with each example stored as an independent model. Similar models are then merged (and thereby generalized); the resulting drop in likelihood is balanced against an increase in the prior. Merging continues until the posterior probability of the model given the data decreases. In the case of probabilistic grammars (Stolcke and Omohundro, 1994), structural priors favor grammars with shorter descriptions, and likelihood is based on the probability of generating the data using the grammar.

We apply a similar strategy to our current task by casting it as a search through the space of possible grammars (or sets of constructions), where the grammars are evaluated using Bayesian criteria. The operations on the set of constructions (merging and composition, described below as **reorganization** processes) extend previous operations to handle relational structures. Similarly, the evaluation criteria need not change significantly for the construction learning case: structural priors favor grammars with fewer, more general constructions that compactly encode seen data; this measure combats the inevitable corresponding drop in the likelihood of generating the seen data using the grammar. Again, the learning algorithm attempts to maximize the posterior probability of the set of constructions given the data.[5]

The main complication requiring a departure from previous work is the need to hypothesize structured maps between form and meaning like those described in the previous section. Essentially, incorporating new data involves both the **analysis** of an utterance according to known constructions and the **hypothesis** of a new construction to account for any new mappings present in the data. These processes, described below, are based on the assumption that the learner expects correlations between what is heard (the utterance) and what is perceived (the situation).[6] Some of these correlations have already been encoded and thus accounted for by previ-

[5]Model merging conducts a best-first search through the hypothesis space based on available merges. It is thus is not guaranteed to find the best model, which would require searching through an exponential number of possible grammars.

[6]The task as defined here casts the learner as primarily comprehending (and not producing) grammatical utterances. The current model does not address production-based means of hypothesizing and reinforcing constructions, which would be included in a more complete model.

ously learned constructions; the tendency to try to account for the remaining ones leads to the formation of new constructions. In other words, what is learned depends directly on what remains to be explained. The identification of the mappings between an utterance and a situation that are predicted by known constructions can be seen as a precursor to language comprehension, in which the same mappings actively evoke meanings not present in the situation. Both require the learner to have an analysis procedure that determines which constructions are potentially relevant, given the utterance, and, by checking their constraints in context, finds the best-fitting subset of those.

Once the predictable mappings have been explained away, the learner must have a procedure for determining which new mappings may best account for new data. The mappings we target here are, as described in the previous section, relational. It is important to note that a relational mapping must hold across arguments that are themselves *constructionally correlated*. That is, mappings between arguments must be in place before higher-order mappings can be acquired. Thus the primary candidates for relational mappings will be relations over elements whose form-meaning mapping has already been established. This requirement may also be viewed as narrowing the search space to those relations that are deemed *relevant* to the current situation, as indicated by their connection to already recognized forms and their mapped meanings.

Details of these procedures are best illustrated by example. Consider the utterance U_1 = "you throw a ball" spoken to a child throwing a ball. The situation S consists of entities S_e and relations S_r; the latter includes role bindings between pairs of entities, as well as attributes of individual entities. In this case, S_e includes the child, the thrown ball and the throwing action, as well as potentially many other entities, such as other objects in the immediate context or the parent making the statement: S_e = {Self,Ball,Block,Throw,Mother,...}. Relational bindings include those encoded by the Throw frame, as well as other properties and relations: S_r = {Throw.thrower:Self, Throw.throwee:Ball, Ball.Color:Yellow, ...}.

In the following sections we describe what the learner might do upon encountering this example, given an existing set of constructions C that has lexical entries for BALL,THROW,BLOCK,YOU,SHE, etc., as well as a two-word THROW-BALL construction associating the `before(throw,ball)` word-order constraint with the binding of Ball to the throwee role of the Throw frame.

Construction analysis and hypothesis

Given this information, the analysis algorithm in Figure 3 first extracts the set F_{known} = {you,throw,ball}, which serves to cue constructions whose form pole includes or may be instantiated by any of these units. In this case, C_{cued} = {YOU,THROW,BALL,THROW-BALL}. Next, the constraints specified by these constructions must be matched against the input utterance and situation. The form constraints for all the lexical constructions are trivially satisfied, and in this case each also happens to map to a meaning element present in S.[7] Checking the form and meaning constraints of the THROW-BALL construction is also trivial: all relations of interest are directly available in the input utterance and situation.[8]

Analyze utterance. Given utterance U in situation S and current constructions C, produce best-fitting analysis A:

1. Extract the set F_{known} of familiar form units from U, and use them to cue the set C_{cued} of constructions.
2. Find the best-fit analysis $A = < C_A, F_A, M_A >$, where C_A is the best-fitting subset of C_{cued} for utterance U in situation S, F_A is the set of form units and relations in U used in C_A, and M_A is the set of meaning elements and bindings in S accounted for by C_A.

   A has associated cost $Cost_A$ providing a quantitative measure of how well A accounts for U in S.
3. Reward constructions in C_A; penalize cued but unused constructions, i.e., those in $C_{cued} \setminus C_A$.

Figure 3: Construction analysis.

In the eventual best-fitting analysis A, the constructions used are C_A ={YOU,THROW,BALL,THROW-BALL}, which cover the forms and form relations in F_A = {you,throw,ball,before(throw,ball)} and map the meanings and meaning relations in M_A = {Self,Throw,Ball,Throw.throwee:Ball}. (Remaining unused in this analysis is the form a.)

We proceed with our example by applying the procedure shown in Figure 4 to hypothesize a new construction. All form relations and meaning bindings, respectively, that are *relevant* to the form and meaning entities involved in the analysis are extracted as, respectively, F_{rel} = {before(you,throw), before(throw,ball), before(you,ball)} and M_{rel} = {Throw.thrower:Self, Throw.throwee:Ball}; the *remainder* of these not used in the analysis are F_{rem} = {before(you,throw), before(you,ball)} and M_{rem} = {Throw.thrower:Self}. The potential construction C_{pot} derived by replacing terms with constructional references is made up of form pole {before(YOU_f,THROW_f),before(YOU_f,BALL_f)} and meaning pole {THROW_m.thrower:YOU_m}. The final

[7] We assume the YOU construction is a context-dependent construction that in this situation maps to the child (Self).

[8] The analysis algorithm can be viewed as a version of parsing allowing both form and meaning constraints. More sophisticated techniques are needed for the many complications that arise in adult language – category constraints on roles may apply only weakly, or may be overridden by the use of metaphor or context. For the cases relevant here, however, we assume that constraints are simple and few enough that exhaustive search should suffice, so we omit the details about how cueing constructions, checking constraints and finding the best-fitting analysis proceed.

construction C_{U_1} is obtained by retaining only those relations in C_{pot} that hold over correlated arguments:

$(\{\texttt{before}(\texttt{YOU}_f,\texttt{THROW}_f)\}, \{\texttt{THROW}_m.\texttt{thrower:YOU}_m\})$

Hypothesize construction. Given analysis A of utterance U in situation S, hypothesize new construction C_U linking correlated but unused form and meaning relations:

1. Find the set F_{rel} of form relations in U that hold between the forms in the analysis F_A, and the set M_{rel} of meaning relations in S that hold between the mapped meaning elements in M_A.
2. Find the set $F_{rem} = F_{rel} \setminus F_A$ of relevant form relations that remain unused in A, and the set $M_{rem} = M_{rel} \setminus M_A$ of relevant meaning relations that remain unmapped in A. Create a potential construction C_{pot} = (F_{rem},M_{rem}), replacing terms with references to constructions in C_A where possible.
3. Create a new construction C_U consisting of pairs of form-meaning relations from C_{pot} whose arguments are constructionally related.
4. Reanalyze utterance using $C \cup \{C_U\}$, producing a new analysis A' with cost $Cost_{A'}$. Incorporate C_U into C if $Cost_A - Cost_{A'} \geq MinImprovement$; else put C_U in pool of potential constructions.
5. If U contains any unknown form units or relations, add (U,S) to the pool of unexplained data.

Figure 4: Construction hypothesis.

At this point, the utility of C_{U_1} can be evaluated by reanalyzing U_1 to ensure a minimum reduction of the analysis cost. As noted in Step 4 of Figure 4, a construction not meeting this criterion is held back from incorporation into C. It is possible, however, that further examples will render it useful, so it is maintained as a candidate construction. Similarly, Step 5 is concerned with maintaining a pool of examples involving unexplained form elements, such as the unfamiliar article a in this example. Further examples involving similar units may together lead to the correct generalization, through the reorganization process to which we now turn.

Reorganizing constructions

The analysis-hypothesis process just described provides the basis for incorporating new examples into the set of constructions. A separate process that takes place in parallel is the data-driven, bottom-up reorganization of the set of constructions based on similarities among and co-occurrences of multiple constructions. Figure 5 gives a high-level description of this process; we refrain from delving into too much detail here, since these processes are closely related to those described for other generalization problems (Stolcke, 1994; Bailey et al., 1997).

Continuing our example, let us assume that the utterance U_2 = "she's throwing a frisbee" is later encountered in conjunction with an appropriate scene, with similar results: in this case, both the unfamiliar inflections and the article are ignored; the meanings are mapped; and constraints with appropriate correlations are found, resulting in the hypothesis of the construction C_{U_2}:

Reorganize constructions. Reorganize C to consolidate similar and co-occurring constructions:

1. Find potential construction pairs to consolidate.
   - **Merge** constructions involving correlated relational mappings over one or more pairs of similar constituents, basing similarity judgments and type generalizations on the conceptual ontology.
   - **Compose** frequently co-occurring constructions with compatible constraints.
2. Evaluate how possible merge/compose operations affect the posterior probability of C on seen data, performing operations on a greedy, best-first basis.

Figure 5: Construction reorganization.

$(\{\texttt{before}(\texttt{SHE}_f,\texttt{THROW}_f)\}, \{\texttt{THROW}_m.\texttt{thrower:SHE}_m\})$

C_{U_1} and C_{U_2} bear some obvious similarities: both constructions involve the same form relations and meaning bindings, which hold of the same constituent construction THROW. Moreover, the other constituent is filled in the two cases by SHE and YOU. As emphasized in our discussion of conceptual representations, a key requirement is that the meaning poles of these two constructions reflect their high degree of similarity.[9] The overall similarity between the two constructions can lead to a merge of the constructional constituents, resulting in the merged construction:

$(\{\texttt{before}(\mathbf{h}_f,\texttt{THROW}_f)\},\{\texttt{THROW}_m.\texttt{thrower:}\mathbf{h}_m\})$

where $\mathbf{h}$ is a variable over a construction constrained to have a *Human* meaning pole (where *Human* is a generalization over the two merged constituents). A similar process, given appropriate data, could produce the generalized mapping:

$(\{\texttt{before}(\texttt{THROW}_f,\mathbf{o}_f)\},\{\texttt{THROW}_m.\texttt{throwee:}\mathbf{o}_m\})$

where $\mathbf{o}$ is constrained to have an Object meaning pole.[10]

Besides merging based on similarity, constructions may also be composed based on co-occurrence. For example, the generalized *Human*-THROW and THROW-*Object* constructions just described are likely to occur in many analyses in which they share the THROW constituent. Since they have compatible constraints in both form and meaning (in the latter case even based on the same conceptual Throw frame), repeated co-occurrence eventually leads to the formation of a larger construction that includes all three constituents:

[9] The precise manner by which this is indicated is not at issue. For instance, a type hierarchy could measure the distance between the two concepts, while a feature-based representation might look for common featural descriptions.

[10] Although not further discussed here, examples with unexplained forms (such as the a in U_1 and U_2) may also undergo merging, leading to the emergence of common meanings.

$$(\{\texttt{before}(\mathbf{h}_f,\texttt{THROW}_f),\texttt{before}(\texttt{THROW}_f,\mathbf{o}_f)\},$$
$$\{\texttt{THROW}_m.\texttt{thrower}{:}\mathbf{h}_m,\texttt{THROW}_m.\texttt{throwee}{:}\mathbf{o}_m\})$$

Note that both generalization operations we describe are, like the hypothesis procedure, merely means of finding potential constructions, and are subject to the evaluation criteria mentioned earlier.

Discussion

We have described a model of the acquisition of grammatical constructions that attempts to capture insights from child language using the formal tools of machine learning. Methods previously applied to word learning are extended to handle grammatical constructions, which are claimed to require new representational and algorithmic machinery.

The model is compatible to the extent possible with evidence from child language acquisition. In particular, the tight integration proposed between comprehension and learning is consistent with usage-based theories of language acquisition: new constructions are hypothesized to capture form-meaning correlations not covered by known constructions, in a manner akin to some of Slobin's (1985) Operating Principles for mapping. The data-driven progression from lexically specific to more abstract constructions is also consistent with Tomasello's (1992) observation that the earliest verb-argument constructions are lexically specific and give way only later to more general argument structure constructions.

More broadly, since the algorithm produces constructions based on any utterance-situation pair and existing set of constructions represented as described above, it can apply equally well for more advanced stages of language development, when the learner has more sophisticated meaning representations and more complex constructions. The potential continuity between early language acquisition and lifelong constructional reorganization offers hope for the modeling of adaptive language understanding systems, human and otherwise.

Acknowledgments

We are grateful for the input and influence of Jerry Feldman and the NTL group at ICSI, as well as comments from the reviewers; opinions and errors remain ours alone. This work was supported in part by an IBM Research Fellowship granted to the first author, and a Praxis XXI Fellowship from the Portuguese "Fundação para a Ciência e a Tecnologia" to the second author.

References

Bailey, D. R., Feldman, J. A., Narayanan, S., and Lakoff, G. (1997). Modeling embodied lexical development. In *Proceedings of the 19th Cognitive Science Society Conference*.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet P roject. In *Proceedings of COLING-ACL*, Montreal, Canada.

Bloom, L. (1973). *One word at a time: the use of single word utterances before syntax*. Mouton & Co., The Hague.

Bloom, P. (2000). *How Children Learn the Meanings of Words*. MIT Press, Cambridge, MA.

Gasser, M. and Colunga, E. (2000). Babies, variables, and relational correlations. In *Proceedings of the Cognitive Science Society Conference*, volume 22, pages 182–187.

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar, Vol. 1*. Stanford University Press.

MacWhinney, B. (1991). *The CHILDES project: Tools for analyzing talk*. Erlbaum, Hillsdale, NJ.

Regier, T. (1996). *The Human Semantic Potential*. MIT Press, Cambridge, MA.

Roy, D. and Pentland, A. (1998). Learning audio-visually grounded words from natural input. In *Proc. AAAI workshop, Grounding Word Meaning*.

Siskind, J. M. (2000). Visual event classification via force dynamics. In *Proc. AAAI-2000*, pages 149–155.

Slobin, D. I. (1985). Crosslinguistic evidence for the language-making capacity. In Slobin, D. I., editor, *The Crosslinguistic Study of Language Acquisition*, volume 2, chapter 15. Erlbaum, NJ.

Stolcke, A. (1994). *Bayesian Learning of Probabilistic Language Models*. PhD thesis, Computer Science Division, University of California at Berkeley.

Stolcke, A. and Omohundro, S. (1994). Inducing probabilistic grammars by Bayesian model merging. In Carrasco, R. C. and Oncina, J., editors, *Grammatical Inference and Applications*, pages 106–118. Springer-Verlag.

Tomasello, M. (1992). *First verbs: A case study of early grammatical development*. Cambridge University Press, Cambridge, UK.

A Model of Infant Causal Perception and its Development

Harold Henry Chaput (chaput@cs.utexas.edu)
Department of Computer Sciences, Taylor Hall 2.124 [C0500]
The University of Texas at Austin
Austin, TX 78712-1188 USA

Leslie B. Cohen (cohen@psy.utexas.edu)
Department of Psychology, Mezes Hall 330 [B3800]
The University of Texas at Austin
Austin, TX 78712 USA

Abstract

The acquisition of infant causal perception has been the center of considerable debate, and some have attributed this phenomenon to an innate causal module. Recent studies, however, suggest that causal knowledge may develop in infants through experience with the environment. We present a computational model of causal knowledge acquisition built using the Constructivist Learning Architecture, a hierarchical self-organizing system. This system does a remarkably good job of developing causal perception from a component view to a holistic view in a way that mirrors data from habituation studies with human infants.

Causal Perception in Infants

Causal perception has been the focus of philosophical inquiry for centuries, but it received its first notable psychological investigation by Michotte (1963). He presented adults with a scene in which one billiard ball struck another stationary ball, resulting in the launching of the stationary ball, and the halting of the moving ball. The subjects, naturally, described this scene as a "causal" event. But by manipulating the launching event (along spatial or temporal dimensions), Michotte could affect a subjects' likeliness of perceiving causality. One can alter the *spatial* component of the event by introducing a gap between the two balls, so that agent and the object never actually touch. Also, one can alter the *temporal* component by introducing a delay between the moment of contact and the moment of launching. As these components deviated from zero gap and zero delay, adult subjects were less likely to classify the event as "causal." These events are illustrated in Figure 1.

Since then, researchers have combined these events with habituation techniques to demonstrate the presence of causal perception in infants. One such study, by Leslie (1984), was able to demonstrate 6 1/2-month-old infants' ability to discriminate between different launching events. Leslie further demonstrated that infants' responses were based, in part, on the causality of the event. For example, infants habituated to a causal event would dishabituate to a non-causal event (e.g. from direct to gap), or vice versa. But infants would not dishabituate to the same degree if the causality remained constant between events (e.g. from delay to delay+gap). Leslie then claimed that these results, since they came from such young infants, were the product of an innate "causal module".

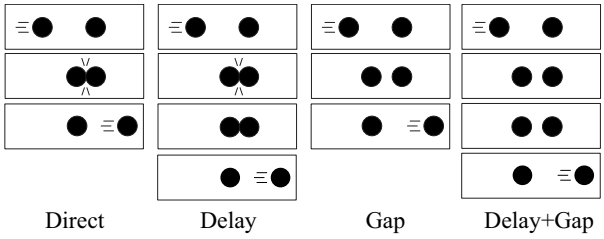


Figure 1: Four different launching events.

More recent studies, though, have cast doubt on a nativist view of causal perception. Cohen and Amsel (1998) performed a similar experiment on 6 1/4-month-old infants, and re-affirmed their ability to discriminate causal events from non-causal events. But they then ran the same experiment on 5 1/2-month-old and 4-month-old infants and found that these younger infants were not able to discriminate between launching events strictly on the basis of causality. Rather, infants responded to the spatial and temporal *components* of the event. Unlike the older infants, younger infants would respond to the introduction or removal or either a delay or a gap, regardless of how this change impacted the causality of the event.

Cohen and Amsel (1998) posited that these results indicated a developmental component to causal perception. It is this progression from component to high-level concept that we are interested in modeling. The development of causality is just one instance of a more general part-to-whole progression that can be seen in a variety of cognitive developmental domains. Cohen (1998) pointed out numerous studies of developmental

cognition that fit this general framework, and proposed an *information processing* approach to cognitive development as an alternative to nativism. Rather than being born with a set of innate perceptual modules, infants start with very low-level perceptual capabilities. This approach is summarized by the following set of propositions (Cohen & Cashon, 2000):

1. Perceptual/cognitive development follows a set of domain-general information processing principles.
2. Information in the environment can be processed at a number of different levels of organization.
3. Higher (more holistic) levels can be defined in terms of the types of relations among lower (parts) levels.
4. Development involves progressing to higher and higher levels.
5. There is a bias to initiate processing at the highest level available.
6. If an information overload occurs (such as when movement is added or when the task involves forming a category), the optimal strategy is to fall back to a lower level of processing.

At the very least, long term development appears to play an important role in the perception of high-level concepts such as causality, regardless of the concept's origin. There are countless learning systems which model knowledge acquisition. But we know of no such model that conforms to the six propositions of developmental information processing described above. There are also very few computational models of infant cognitive development that differentiate between long-term learning and short-term habituation, let alone use one to determine the other. One example in the language development domain has recently reported by Schafer and Mareschal (2001).

Constructivist Learning Architecture

The Constructivist Learning Architecture (CLA) (Chaput, 2001) is a hierarchical self-organizing system designed to generate concepts at multiple levels of abstraction through the organization of sensory input. Using the six propositions of cognitive development listed above as a design specification, CLA was built to learn hierarchical knowledge structures through observation, and use those structures to produce the kind of short-term habituation effects that infants exhibit throughout their development.

The information processing approach to cognitive development described above does not mention any kind of corrective feedback, and thus suggests an unsupervised learning system. For this reason, CLA uses one such system, the Self-Organizing Map (Kohonen, 1997), as a building block. The Self-Organizing Map (SOM) is a two-dimensional matrix of nodes, each of which stores a feature vector. As stimuli are repeatedly presented to the SOM (as feature vectors), the SOM adjusts the feature vectors of its nodes to represent the observed stimuli. A trained SOM exhibits the following attributes: 1) the feature vectors of nodes in a SOM reflect the stimuli presented to the SOM (*environmental representation*); 2) nodes which are close to each other in the network have similar feature vectors (*similarity topography*); and 3) stimuli which occur more often will be represented by a larger number of nodes (*frequency effect*).

But although the SOM performs the kind of unsupervised category formation that appears to be at work in cognitive development, it does not by itself form the kind of hierarchical knowledge representation suggested by the information processing approach.

CLA achieves this hierarchical representation by connecting multiple SOMs into a hierarchy. Like a regular SOM, the lowest layer of CLA (Level 1) receives raw input from the environment. When a stimulus is introduced, each node in the Level 1 layer receives an activation, A, which is in proportion to how close the stimulus is to the nodes' representation. (This is determined using a Euclidean distance metric.) These activation values are then collected for the layer into a corresponding matrix of activation values, or an *activation matrix*. This activation matrix then becomes the input vector to the layer directly above. This process then repeats for as many layers as are contained in the whole system. For an illustration, see Figure 2.

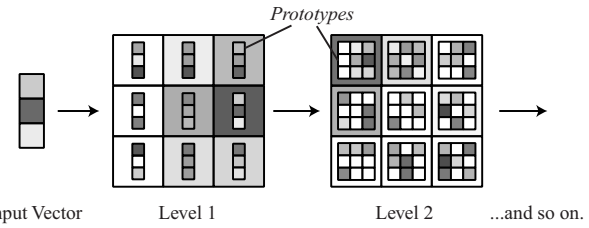


Figure 2: The first two layers of an example CLA. The darkness of each cell represents its level of activation.

Of course, two SOMs connected in this fashion can learn no more or less than a single SOM would. But the stratification of the SOMs allows for processing to occur at intermediate stages. There are three types of intermediate processing that are relevant to the present discussion: activation decay, activation blurring, and multi-modal layers. *Activation Decay* allows the activation of a given node to decay over time, rather than reset at each time step. This is useful as a rudimentary representation of time and sequence. *Activation Blurring* will "blur" the activation matrix by applying a Gaussian filter to the matrix. This has the effect of spreading the activation of a given node to

surrounding nodes, which is particularly useful given the SOM's similarity topography feature. Finally, a layer can receive input from two or more lower-level layers. We call these layers *Multi-Modal Layers*.

The result is a system that will have the information processing properties listed above, which we address one by one. First, CLA is not designed for any particular domain, but is a domain-general learning architecture that can readily be applied to any number of areas in cognitive development. Second, when a stimulus is introduced to the system, it will create activation patterns at multiple layers, which are actually different levels of organization; processing of these patterns can occur at any of these layers. Third, because each layer is organizing activation patterns of the layer beneath it, higher-level representations can be defined in terms of the types of relations among lower-level representations.

Fourth, learning in CLA involves progressing to higher and higher levels. When CLA is learning, each layer is trying to build categories out of the activation patterns of the layer beneath it. While one layer is organizing, all the layers above it cannot form stable categories because the underlying activation patterns are unorganized. Only when a layer "settles" into a coherent organization can the layers above it organize. The result is a progression from level to level.

Propositions 5 and 6 involve the resulting behavior of the system, rather than its organization. This paper does not address these propositions directly, but we do consider their ramifications in the discussion section.

Experiment: Learning a Causal Event

We conducted an experiment designed to show whether CLA could model the data produced by human infant subjects in the Cohen and Amsel (1998) study, given the same experiment design and starting set of assumptions. Specifically, we are looking to see if our model exhibits the part-to-whole progression demonstrated in infants between 4 and 6.25 months.

Design

Cohen and Amsel (1998) posit that infants are able to perceive the spatial and temporal components of a launching event before they perceive its causality. For this reason, we present the launching events to the learning system by means of two input vectors. The first input vector (the "movement vector") reported the magnitude of the movement of each of the two balls with two rows of nodes. By ignoring the position of each ball, this layer would use activation decay to represent the temporal information of the launching event and exclude the spatial information. The movement vector represented the movement of each ball, with each ball represented in its own row. The element of the row corresponding to the amount of absolute change in position, scaled to the width of the input vector, was set to 1.0. So, elements ~~to~~ on the right side of the vector represent rapid movement, while the left most element would represent no movement. These values decay over time.

Figure 3, for example, shows the state where the first ball (represented in the top row) is now stationary, but was recently moving; while the second ball (represented in the bottom row) is now moving, but was recently stationary. This state occurs in a direct launching event shortly after a collision.

Figure 3: The movement input vector. The decay is shown by a decrease in brightness (or change in color from yellow to red).

The second input vector (the "position vector") reported the position of the balls on the table. The position was represented by a 20 element vector. (The vector only had one row of nodes because the collisions presented were always horizontal.) The positions of the balls on the table were scaled to the 20 element vector, and the element nearest to each ball's position was set to 1.0. Complimentary to the movement vector, the position vector reports the spatial information and excludes the temporal information.

Figure 4: The position input vector.

In Figure four, we see the state where the two balls are close, but not touching. Like Figure 3, this state occurs in a launching event shortly after a collision.

Receiving each input vector was a 5-by-5 node layer to observe it. The "movement layer" organized activation patterns in the movement vector, while the "position layer" organized activation patterns in the position vector. These Level 1 layers would learn the independent spatial and temporal features of the launching event.

Finally, there was a Level 2 layer (the "Top Layer"), which observed *both* bottom-level layers. This layer, 6-by-6 nodes, would learn the combination of activation patterns in the Level 1 layers. Thus, it should discover the combination of spatio-temporal events that comprise each of the events.

The movement vector had an activation decay of 0.25, so that the movement layer could learn the temporal attributes of the ball movements. All other layers had an activation decay of 1.0 (instant decay). Similarly, the position vector had its blurring set to 0.6 to let it see proximal ball positions as similar (that is, having a ball at position 3 is very similar to having a ball at position 4, but very different from having a ball at position 17). All other layers had their blurring set to 0.0 (that is, turned off).

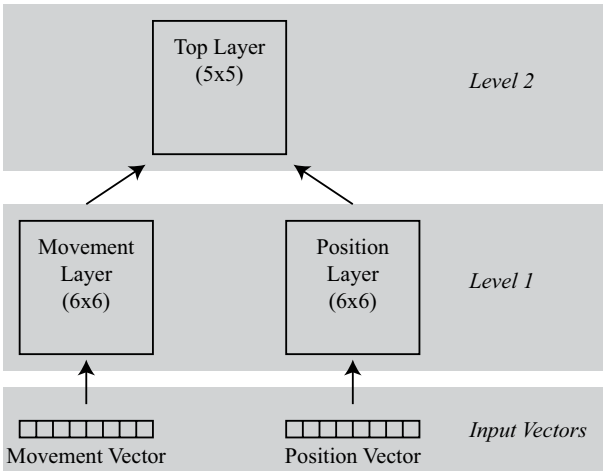


Figure 5: A schematic of the CLA used in the causality experiment.

To generate the launching events, we used a simulated "pool table" written for this experiment. This pool table is used for both the long-term training of the learning system as well as the short-term habituation studies. A simple algorithm was used to represent the state of the pool table through the two input vectors.

For long-term training (meant to represent an infant's long-term experience with the world), the learning system was presented with 2500 complete launching events. The type of launching event presented to the learning system was chosen using a probability meant to approximate roughly the nature of the real world: a direct launching event had a 0.85 probability of being chosen, while delay, gap, and delay+gap events each had a probability of 0.05. The presentation of a complete launching event constituted a single training cycle. The learning rates and neighborhoods of each SOM in the CLA system were decreased as training progressed. (Learning rate decreased from 0.1 to 0.0 over 1000 cycles, and the neighborhood from 1.0 to 0.0 in the same time frame.) This is the customary training procedure when using SOMs.

In order to simulate the changes during the short-term habituation trials, a "familiarity" variable F was associated with each node. Remember that we associate an activation A with each node. Familiarity for each node always approached the node's activation by some rate using the formula $F=F+r(A-F)$, where r is the rate of approach. We then determined an output O for each node using the formula $O=A^{(1.0-(A-F))}$. Recalling that both F and A are between 0.0 and 1.0, the result is that output levels are amplified, relative to raw activation, as activation differs from the familiarized level. (See Figure 6 for a graphical representation of these values.) The output for each node in a layer was summed to create a Layer Output. This was then averaged across the duration of the event, giving us a Mean Layer Output. Dishabituation was measured as *change* in Mean Layer Output.

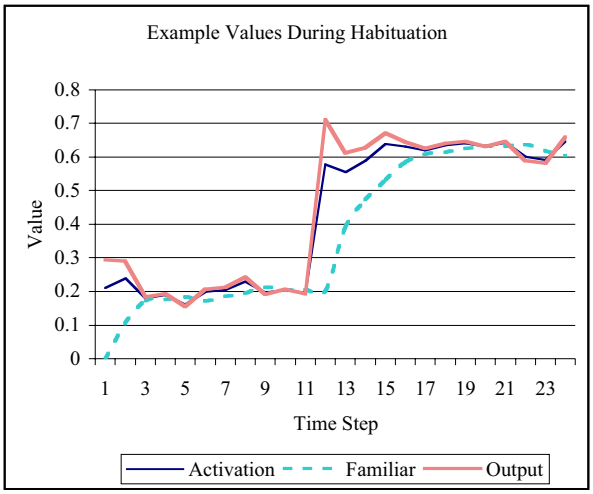


Figure 6: Example values of Activation, Familiarization and Output for a single node during the habituation trials. An event change at time step 12 is represented by a jump in activation.

For each habituation trial, the network was exposed to five repetitions of the habituation event, and then exposed to the test event. A complete habituation experiment consists of 16 parts: four habituation events by four test events. Familiarity levels were cleared for all nodes before each part of the habituation experiment. In all, 12 "simulated infants" were fully trained and tested.

Results

Because of the nature of the events we are dealing with, the difference between a delay event and a gap event should be the same as the difference between a direct event and a delay+gap event. This is because both pairs involve equal changes along the spatial *and* temporal axes. (See Figure 7.)

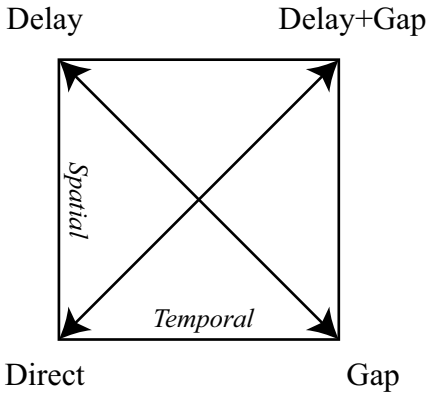


Figure 7: Launching events along spatial and temporal axes. Events at opposite corners involve both a spatial and temporal change of equal amounts and, thus, should be equivalent given a component model.

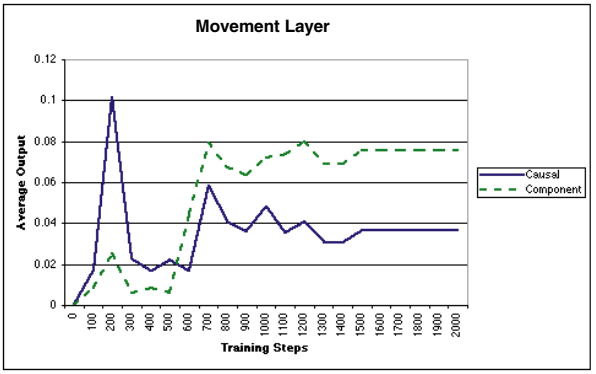
We averaged the dishabituation levels for delay-to-gap trials with gap-to-delay trials, and compared them to the average of direct-to-delay+gap trials and delay+gap-to-direct trials. We did an analysis of variance on both of the Level 1 layers and found that, in fact, both showed a significantly greater response to delay-gap changes, $F(1,11) = 243.3$, $p < .0001$ and $F(1,11) = 34.4$, $p < .0001$ for the movement and position layers, respectively.

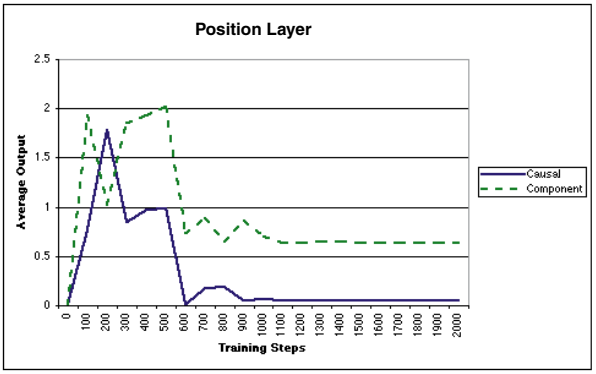
This odd disparity comes about because we have designed each of the Level 1 layers to train on one component exclusively, to the exclusion of the other. For example, the movement layer, which is sensitive to temporal differences, is *insensitive* to spatial differences, so that a direct event and a gap event look nearly identical. For this reason, we would expect this layer to see the direct event and the delay+gap event as similar, since the direct event looks like a gap event, and the delay+gap event also has a gap. The converse is true for the position layer. Thus, we would expect these two layers to respond more to the delay-to-gap change.

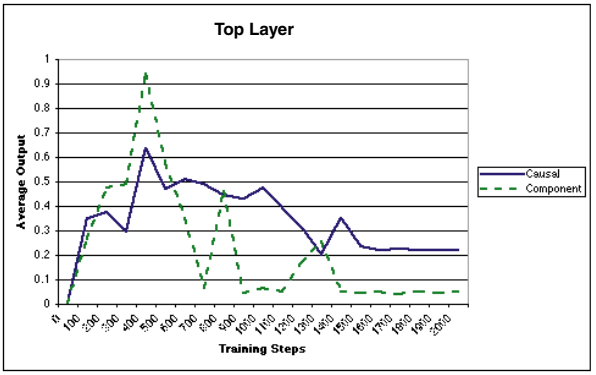
We verified that this difference was the result of the exclusivity of the input by comparing the average of trials where there was strictly a spatial change to the average of trials where there was strictly a temporal change. There was a significantly greater response to temporal changes than spatial changes in the movement layer, $F(2,22) = 4.1$, $p < .05$. And, conversely, there was a significantly greater response to spatial changes in the position layer, $F(2,22) = 123.2$, $p < .0001$.

Having verified that our Level 1 layers were operating according to our expectations, we then wanted to see if the Top Layer was responding to the components of the events or to their causality. We ran an analysis of variance on the same two event types as above: direct-delay+gap and delay-gap. Without the component exclusivity present in the Level 1 layers, the difference between these two conditions should be the same. However, there was a significantly greater response to the direct-delay+gap change than to the delay–gap change, $F(2,22) = 15.3$, $p < .0001$. This shows a clear preference on the basis of causality rather than just the independent components.

We can see, too, that this difference in processing has a developmental or long-term experiential component. Figure 8a and 8b shows the preference for a component model of causality settling in the two Level 1 layers after about 800 training cycles. Figure 8c shows that the Top Layer does not settle on a causal model until about 1500 cycles. As mentioned earlier, this is because of the nature of CLA: the lower levels must settle before the higher levels can.







Figures 8a, 8b and 8c: The acquisition of a component model vs. a causal model in different layers over time.

Discussion

Our CLA system has created a hierarchical knowledge structure that can produce habituation results compatible with those from similar studies with human infants. These results are consistent with those of Cohen and Amsel (1998) as well as Leslie (1986). Contrary to Leslie's conclusions, however, our model does not rely on a causality module.

That is not to say, of course, that our model has *nothing* built in to it. CLA is not *tabula rasa.* Unlike a modularist view, though, the innate attributes of CLA are domain general information processing principles. More generally, CLA has innate processes, rather than specific innate knowledge of abstract concepts.

CLA also relies on the vast majority of direct events in the world compared to non-direct events. We believe that infants also rely on this arrangement. CLA is guided by the nature of the environment to develop a causal model because there are simply more direct events in the environment. We can imagine an alternate universe which contained more delay events than any other kind, and our model would predict that development in this kind of environment would result in a drastically different world view.

One might ask what the point of having a stratified representation of causality might be, when it might be possible to achieve this same learning with a monolithic system. As previously stated, our hierarchical approach has the effect of producing the stages in development that we see in infants. But more than just fitting the experimental data, a hierarchical representation makes it possible to address the last two information processing principles described above. Cohen and Cashon (2000), and others, have observed hierarchical knowledge processing in infants, both in terms of perceptual preference and in handling cognitive overload. CLA's hierarchical design makes such processing possible, where a monolithic system would not. We intend to use CLA for robotic control, and we feel that principles five and six can be used with CLA's knowledge hierarchy to give certain layers priority over others. Also, we plan to test proposition six by overloading our system and seeing if it produces the "fall back" phenomenon that has been demonstrated in infants.

Conclusion

Although there are several connectionist models of infant development, CLA is the first to use hierarchical representation and differentiate between long-term and short-experience. These are important factors in cognitive development, and are often not given much weight even in *real* infant habituation experiments. The information processing approach to cognitive development has been applied to infant cognition with considerable success. We feel that a computational model which uses this approach holds promise for modeling the acquisition of a variety of domains within infant, child, and even adult cognition.

Acknowledgments

This research was supported in part by NIH grant HD-23397 to the first author from the National Institute of Child Health and Human Development.

We would like to thank Risto Miikkulainen for his advice and guidance at various stages of this project.

References

Chaput, H. H. (2001). Post-Piagetian constructivism for grounded knowledge acquisition. To appear in *Proceedings of the AAAI Spring Symposium on Learning Grounded Representations*, March 2001, Palo Alto, CA.

Cohen, L. B. (1998). An information processing approach to infant perception and cognition. In G. Butterworth and F. Simion (Eds.), *Development of Sensory, Motor, and Cognitive Capacities in Early Infancy: From Sensation to Cognition.* Sussex: Erlbaum (UK) Taylor & Francis.

Cohen, L. B. & Amsel, G. (1998). Precursors to infants' perception of the causality of a simple event. *Infant Behavior and Development 21 (4)*, 713-732.

Cohen, L. B., Amsel, G., Redford, M. A. & Casasola, M. (1998). The development of infant causal perception. In A. Slator (Ed.), *Perceptual development: Visual, auditory and speech perception in infancy*. London: UCL Press (Univ. College London) and Taylor and Francis.

Cohen, L. B. & Cashon, C. H. (2000). Infant object segregation implies information integration. *Journal of Experimental Child Psychology*, (in press).

Cohen, L. B. & Oakes, L. M. (1993). How infants perceive a simple causal event. *Developmental Psychology,* Vol. 29, No. 3, 421-433.

Kohonen, T. (1997). *Self-Organizing Maps*. Berlin: Springer-Verlag.

Leslie, A. M. (1986). Spatiotemporal continuity and the perception of causality in infants. *Perception, 13,* 287-305.

Michotte, A. (1963). *The Perception of Causality*. New York: Basic Books.

Schafer, G. & Mareschal, D. (2001). Modeling infant speech sound discrimination using simple associative networks. *Infancy 2*, 7-28.

The Effect of Practice on Strategy Change

Suzanne C. Charman (CharmanSC1@cardiff.ac.uk)
School of Psychology, Cardiff University, Cardiff CF10 3YG, Wales, United Kingdom

Andrew Howes (HowesA@cardiff.ac.uk)
School of Psychology, Cardiff University, Cardiff CF10 3YG, Wales, United Kingdom

Abstract

Some effort has been made to examine why people often do not adopt efficient strategies, however little attempt has been made to determine the conditions that support strategy generation. This study examines how practice may lead to efficient strategy generation, through perceptual noticing and the elaboration of a device representation. Forty-three participants were required to complete drawing tasks in MS PowerPoint, for which there are a number of possible strategies that share the same component operators, and yet vary in efficiency. Merely by practicing a component of a less efficient strategy, a more efficient strategy was generated earlier on in the experiment. Further, the efficiency of the strategy used at test was correlated with device knowledge. It is proposed that through practice a user's device representation becomes elaborated, and this in turn leads to strategy generation. The possibility of a perceptual noticing mechanism for problem solving was also investigated, however providing strong perceptual groupings did not aid strategy generation.

Introduction

A reliable finding in experimental psychology is that practice on a task leads to faster and less erroneous performance of that task. In the 1920s Snoddy (1926, as cited in Anderson, 1995) graphed learning data and demonstrated that a power function characterises the rate of improvement in the performance of a task. This means that with practice the speed at which people complete a task increases with diminishing returns (Newell and Rosenbloom, 1981).

A growing body of evidence suggests that power law learning occurs within strategy, and not with regards to the task as a whole (Rickard, 1997; Delaney, Reder, Staszewski and Ritter, 1998). Compton and Logan (1991) suggest that improvement in the performance speed of a task is often due to the adoption of more refined and effective strategies.

However, there appears to be evidence that people do not adopt efficient strategies as rapidly as might be expected. Early explanations for this failure centred around the Einstellung effect, where prior experience results in a reluctance to investigate alternative procedures for a novel task (Luchins, 1942). Inefficient use of a device was also investigated using the ACT theory of skill acquisition (Anderson, 1982) where procedures that hinder completion of a task are discarded upon detection, however procedures that are sufficient and yet inefficient are less easily identified and so are maintained. In considering the problem of inefficient device use, Carroll and Rosson (1987) suggest that people are trapped by a production paradox, meaning that users are so focused on completing tasks, that they are unlikely to take time out to learn about a device. Paradoxically, if time were spent learning about the device, performance might be improved in the long term.

More recent evidence suggests that even people who are very skilled at a set of tasks are also not likely to operate as efficiently as might be expected (Nilsen et al., 1993; Nilsen, Jong, Olson and Polson, 1992). For example Young and MacLean (1988) found users do not always choose the faster method when presented with several different routines. Bhavnani and John (1997) observed that even after several years of experience and formal training in a particular CAD package, many users had not adopted efficient strategies. Further, the use of inefficient strategies impacted upon completion times and caused an increase in low level error. Bhavnani, John and Flemming (1999) highlight the difficulty of moving from "sufficient use to a more efficient use of computer applications" (p.183). The reason for this inefficiency, they suggest, is not related to the standard of interface design or experience with the package, but to strategic knowledge. Once participants received both traditional command based training ('learning to do') and strategic knowledge training ('learning to see' or recognize opportunities for efficient strategies to be used), it was found that most tasks were performed using efficient strategies (Bhavnani et al., 1999; Bhavnani, in press). However, the relative success of those who received strategic knowledge training should not be too surprising, as during the extensive training stages participants were explicitly taught each strategy. The tests therefore are more of an ability to recall and apply efficient strategies to novel situations, rather than strategy generation per se.

Some would argue that the use of strategies, that on the surface appear inefficient, could actually be rational.

Potential costs associated with strategy generation, such as time spent exploring the device and searching through problem space, could outweigh the benefits that a possible new strategy may deliver. It may therefore be rational to maintain sufficient, yet inefficient procedures. For example with regards to choice between strategies, Young and MacLean (1988) found that where the effort of using a strategy is perceived to be high, people are prepared to 'trade off' the possible benefits of using that strategy. Users choose to continue with a method that is easier to implement, yet results in a slower task completion time.

Early attempts at investigating strategy change by Anzai and Simon (1979) found that participants spontaneously advanced through several strategies when solving the Tower of Hanoi task. Their explanations for efficiency gaining strategy changes included both mechanisms that perform modifications on existing procedures to make them more efficient (such as the identification and removal of redundant steps), and also a mechanism for perceptual noticing. In the Tower of Hanoi, this perceptual noticing mechanism identifies that the problem can be restructured in terms of pyramids.

Recent attempts to model strategy generation and change include Shrager and Siegler's (1998) SCADS simulation. Mechanisms for strategy generation and change proposed by Crowley, Shrager and Siegler (1997) were used to simulate children's use of addition strategies. Shrager and Siegler's (1998) model represents a significant step forward in understanding. However, due to the focus of the model, the reasons that people fail to apply the hypothesised mechanisms for strategy generation are not considered.

The study reported provides evidence for the conditions that support strategy generation. While some effort has been spent examining the failure of people to use optimal strategies, little evidence exists about how and when people generate new strategies. Crowley et al. (1997) suggest that during practice people make a method more efficient by identifying and eliminating redundant sub-procedures. However it is possible that practice results in efficient performance through other mechanisms such as perceptual noticing of task related features and elaboration of the user's device representation. Perceptual noticing may involve both noticing the structure of the task and the structure of objects to which operators are applied. Known operators could, as a result, be applied more efficiently. Practice may also allow elaboration of the device model through incidental learning, and this in turn may permit a more comprehensive, high quality search for efficient procedures.

The task used in the experiment is a computer-based drawing task, similar to that used by Bhavnani et al. (1999), and was selected for its ecological validity. The task provides a rich domain for studying strategy change as the basic fence, copy and paste operations can be embedded within a range of relatively sophisticated strategies for complex drawings. For example, people asked to draw a series of repeated items may use the recursive multiple-item-copy (MIC) strategy. This is where a single item is copied, then the resulting pair, and then all of the objects are copied again. The number of objects increases exponentially.

Also, following the proposal of a perceptual noticing mechanism by Anzai and Simon (1979), the task was either presented with a strong pattern that contained groupings useful to the development of the efficient strategy, or with no groupings. A mechanism for perceptual noticing should lead the user to be more likely to generate a multiple item copy strategy where groupings allude to the strategy.

The study is designed to determine whether merely practicing known procedures makes the generation of efficient strategies more likely. The importance of a mental representation of both the device and task, and their development through practice, will be examined. It is hypothesised that practice on the components of a non-optimal strategy can establish the prerequisites, through elaboration of the device representation, for the generation of a new strategy.

Method

Participants

Forty-three regular computer users, first and second year psychology undergraduates ranging in age from 18 to 32, took part in the experiment for course credit. All participants were given 2 hours of course credits to take part in the study, no matter how long they took (the average time taken was approximately 1 1/2 hours), in order to encourage efficient completion of the tasks.

Design

The study involved two between subjects manipulations. Although all of the participants knew the component procedures necessary for the efficient strategy to be used, the manner in which they were practiced varied. Practice trials involved different objects, same objects or same objects with space between them (see Table 1). The second manipulation was the pattern of the test item. In order to use MIC it was hypothesized that a participant must make certain perceptual groupings. Where there was a patterned test item, the objects were arranged so that the groupings necessary to use MIC were already present (Figure 1).

Procedure and Materials

The participants completed an informed consent form and a brief questionnaire to determine prior experience

with Microsoft PowerPoint, as well as other software packages with drawing functions. The tuition phase was then completed, which ensured that the participants mastered basic drawing skills (such as drawing, moving and altering shapes, and selection of single items by fencing), and were also made aware of the existence of some functions, including copy and paste. The participants were informed that they should only use functions identified in the tutorial stage. These included fencing, copying and pasting, but, for example, excluded duplication and grouping.

After the tuition phase the participants completed an open-ended questionnaire designed to assess knowledge about the device. Ten questions relevant to the key concepts particular to the MIC strategy were included. Five questions related to fencing multiple objects with space between them and five related to the manipulation of multiple objects. The participants were then given the same pre-test version of the test stimuli and asked to complete the drawing in as few moves as possible. The pre-test item consisted of eight equally spaced two-item objects (P- in Figure 1). If a participant completed the pre-test using any form of MIC (see Table 2) they were excluded from the analysis.

The main series of items were then presented. Each participant carried out four practice trials (see Table 1), filled in the device representation questionnaire (same as before), and then completed a saturation trial. The saturation trial involved drawing one shape, fencing it and using copy and paste to create another identical shape. This was designed to make sure all participants had fence, copy and paste functions readily available to them in memory. Half of the participants were then presented with a patterned test trial (P+) and half were presented with a non-patterned test trial (P-).

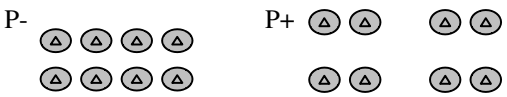
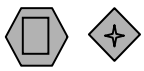
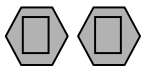


Figure 1: Non-patterned and patterned test trials

Four practice trials, the device representation questionnaire, saturation trial and test stage were repeated four times. All participants received exactly the same instructions and drew the same stimuli, except for the type of practice trials and the pattern of the test trials. Therefore each group had the same opportunity (in five test trials) to use the efficient strategy.

Table 1 shows examples of the practice tasks. The different practice trials allowed differential experience with the operators fence, copy and paste. All the participants used the operators in the saturation trial, yet in the practice trials they were used in different ways.

Table 1: Practice trials

Practice group	Explanation
1: Different objects 	Participants drew each shape one-by-one, as it was not possible to use fence, copy and paste.
2: Same objects 	The participants drew the first two shapes that constitute the first object, and used fence, copy and paste to complete the task.
3: Same objects with space 	Participants drew the two shapes on the left and used fence, copy and paste to complete the drawing.

If participants had not used the exponential MIC strategy during the first five test trials they were given five more opportunities to do so. They were instructed to complete the task as efficiently as they could, minimising the number of steps taken to complete the task.

Microsoft PowerPoint '97 was used to carry out the drawing tasks, these were all video recorded.

Coding of Strategies

There are at least seven strategies that can be used to complete the task with the functions made available to the participants (those identified in the tuition phase). Multiple Items Copy (MIC) is the most efficient manner of performing the task. This strategy involves fencing (as depicted in Figure 2), where all objects within the fence become selected, and the manipulation of more than one object at a time.

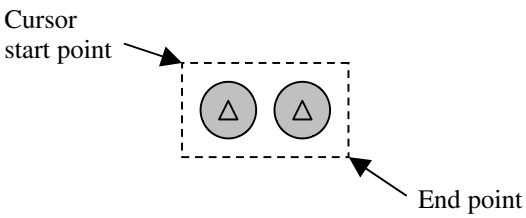


Figure 2: Fencing, by using the mouse to click and drag from the start point and releasing at the end point

There are some key concepts that must be understood before each strategy (see Table 2) can be used. Each practice group differs in their experience of these concepts. Firstly it must be appreciated that copy and paste can be used on a single item once it is selected, this being a central concept for the use of Element Copy strategy. All practice groups experience this through the saturation trial. Secondly it must be understood (to use DAC and MIC) that more than one item may be

selected by using a fence, and that copy and paste can be performed on all selected items at once. Only the two 'same objects' practice groups (2 & 3) experienced this. Finally, to use MIC it must be appreciated that items with space between them can be selected by using a fence, and that all selected items can then be manipulated together. Only the 'same objects with space' practice group experienced this.

Table 2: Possible strategies for completion of the task, result of GOMS analysis and points awarded

Strategy name, points for use and GOMS analysis result	Classification requirements and Example 
Element-by-element (EBE) 1 *153s*	Each two item picture is drawn element by element (a square would be drawn and then a triangle, process repeated 7 times).
Division (D) 2 *133s*	All of the first shape are drawn and then all of the other shape (all the squares drawn, then all triangles).
Element copy (EC) 3 *101s*	Copy and paste are used on single shapes (one square would be drawn, copied and pasted 7 times, and then the same for the triangle).
Detail Aggregate Copy (DAC) 4 *64s*	All the details are completed in the first object, then it is fenced, copied and pasted seven times (a house would be drawn and then fenced, copied and pasted 7 times).
Multiple Items Copy (MIC_1) 5 *53s*	As with DAC, but once the first 4 copies are in place, they are all fenced, copied and pasted to make 8.
Multiple Items Copy (MIC_2) 6 *50s*	As with DAC, but once the 2nd copy is in place, both are fenced, copied and pasted to make 4, pasted again to make 6, and pasted again to make 8.
Exponential MIC (MIC_{exp}) 7 *47s*	As with DAC, but once the second copy is in place both are fenced, copied and pasted to make 4. The 4 are then fenced, copied and pasted to make 8.

A GOMS analysis was carried out to determine the efficiency of each strategy. Points for each strategy are allocated on the basis of this analysis (see Table 2). Where available the moves were assigned the length of time specified by Olson and Olson (1990). Times for moves not covered in the literature were determined from an observational pilot study. The seven strategies were classified on a seven-point scale with 1 being the least efficient and 7 being the most efficient strategy.

Results

Seven of the forty-three participants used a form of MIC during the pre-test stage, and so were excluded from the analysis, leaving thirty-six. All participants were experienced users of at least one Microsoft package, yet inexperienced with Microsoft PowerPoint.

At the pre-test stage of the experiment there were no between group differences in time taken to perform the task, strategy used and device knowledge. An overall speed up in the performance of the task was observed, and by the fifth trial a main effect of practice on completion time was approaching significance [$F(2,30)=3.082$, $p=0.06$, $MSE=405.3$] (different objects $M=69$s, same objects $M=56$s, and same objects with space $M=49$s,).

Best Strategy Used at Test

Participants were given a strategy score ranging from one (inefficient strategy) to seven (efficient strategy) for each of the five test trials (see Table 2), and for the second set of test trials undertaken at the end of the experiment (in the event that MIC was not used earlier). A between subjects three by two ANOVA found a significant main effect of practice on the best strategy used at test [$F(2,30)=7.784$, $p<0.01$, $MSE=1.6$]. No main effect of pattern was found.

A Tukey HSD test confirmed a significant difference between the different ($M=4.4$) and same ($M=5.9$) objects conditions ($p<0.05$) and between the different and same objects with space ($M=6.3$) conditions ($p<0.01$). The difference between the two same objects conditions did not reach significance. The same pattern of significant results was found when considering the best strategy used over all ten test trials.

Strategy Use Score

For each participant the sum of strategy scores over the five test trials was taken as the strategy use score, and was essentially a measure of overall efficiency. A two by three between subjects ANOVA was performed on the data, and as before a significant main effect of practice was found [$F(2,30)=6.405$, $p<0.01$, $MSE=20.6$].

A Tukey test confirmed a significant difference between the different ($M=20.2$) and same objects ($M=25.0$) conditions ($p<0.05$) and between the different and same objects with space ($M=26.5$) conditions ($p<0.01$). The two same objects conditions were not significantly different.

Discovery Trial

A significant main effect of practice (but not pattern) was found for the trial (1-10) upon which one form of MIC was first used by a participant (those that did not use MIC at any time in the experiment were given a

score of 10), $F(2,30)=13.826$, $p<0.001$, $MSE=3.7$. A Tukey test found significant differences between the different ($M=6.9$) and same objects ($M=3.5$) conditions ($p<0.001$), and between the different and same objects with space ($M=3.2$) conditions ($p<0.001$).

There was no significant main effect of pattern, although a significant interaction (shown in Table 3) was found [$F(2,30)=6.337$, $p<0.01$] between pattern and practice. Simple effects tests found practice had a significant effect where the test item was patterned [$F_A@b1(2,30)=19.400$, $p<0.01$], and pattern had a significant effect where practice trials involved drawing different objects [$F_B@a1(1,30)= 8.046$, $p<0.01$].

Table 3: The mean trial upon which MIC was first used

	Practice		
	Different Objects	*Same Objects*	*Same Objects with Space*
Patterned	8.5	2.8	2.2
Non-patterned	5.3	4.2	4.2

Device Representation and Strategy Generation

The device representation questionnaire (DRQ) provided a score, out of ten, that reflected the knowledge each participant had about the device. This measure was repeated throughout the experiment and was specific to aspects of the device central to a MIC strategy. A Spearman's non-parametric correlation between improvement in the performance of the task (difference in strategy score from test trial one to test trial five) and the improvement in DRQ score was significant ($r_s=0.384$, $p<0.05$).

A two by three between subjects ANOVA was conducted on the score for each DRQ. Before practice, scores on the DRQ did not differ between groups. Results for all DRQs administered after practice (DRQs 2-5) followed the same pattern, and so the scores were combined. No main effects of pattern and practice on DRQ score were found, however interactions between practice and pattern were significant [$F(2,30)=7.312$, $p<0.005$, $MSE=3.3$] (see Table 4).

Table 4: Average score for DRQs 2-5

	Practice		
	Different Objects	*Same Objects*	*Same Objects with Space*
Patterned	5.8	7.9	8.9
Non-patterned	8.2	6.1	5.6

Simple effects tests revealed that practice had an effect where the test item was patterned [$F_A@b1(2,30)= 4.793$, $p<0.05$]. Pattern was found to have an effect on device representation where different objects were drawn at practice [$F_B@a1(1,30)=5.340$, $p<0.05$] and where objects drawn at practice were the same with space between them [$F_B@a3(1,30)=8.005$, $p<0.01$].

The questionnaire measured understanding of two concepts central to the use of MIC. Firstly, that multiple objects with space between them can be selected at the same time by using a fence, and secondly that multiple objects can be manipulated simultaneously once selected. For each of these concepts there were five questions. The trial upon which participants reached a good understanding of these concepts was taken to be when they answered four or all five of these questions correctly. An ANOVA revealed an interaction between practice and pattern for the trial upon which participants reached a good understanding of both fencing [$F(2,30)=3.451$, $p<0.05$, $MSE=1.8$] and manipulating objects [$F(2,30)=6.843$, $p<0.005$, $MSE=1.1$]. Simple effects tests found the same pattern of significant results for manipulating objects as were found on the overall DRQ results (Table 4). Further analyses found that a good understanding of fencing was reached significantly earlier on than a good understanding of manipulating objects [$t(35)=7.402$, $p<0.001$].

Discussion

As expected, practice resulted in the more efficient performance of the task. Those in groups where practice involved the selection and manipulation of more than one item performed the task more efficiently overall, generated more efficient strategies, and did so earlier on in the experiment. Although all groups used fence, copy and paste, the manner in which they were experienced influenced strategy generation. Most importantly, repeated use of the component parts of the less superior strategy proved useful for the generation of the new strategy (MIC). In addition, the more participants understood about the device, the better the strategy that was used at test.

For the copy paste task described, the results imply that people acquired information about the device through the repeated practice of a known method. Initially a good understanding of the fence operator was gained, and then an elaborated model of manipulating multiple objects. Participants learned more than was required to merely reduce the performance time of the method, as the knowledge acquired correlated with the efficiency of the strategy generated. This finding has implications for models of practice that assume people merely re-code previously stored information. Neither Anderson's compilation (1982), nor Rosenbloom and Newell's (1986) chunking model predict the results reported here. Both models explain how an existing strategy becomes more efficient, and thus effect a speed-up, rather than how information, acquired through practice, supports strategy change. Logan's (1988) instance-based model also fails to predict the

results reported here, as an increase in the number of instances of a strategy in memory can only support shift from algorithm to memory-based processing, not the kind of strategy change reported here. The results also have implications for models of strategy change. In Shrager and Siegler's (1998) model the role of new information acquired through repeated practice is not considered as a precursor to strategy change.

Also in contrast to Anzai and Simon (1979), providing the perceptual groupings that must be made in order to use the efficient strategy had no effect on strategy generation. However, groupings did have an effect on strategy generation depending upon the practice experienced. An explanation for this interaction may be that those in the different objects group, where efficient strategies were not readily available, had a relatively high workload. Constructing the pattern may have added to the high workload of the group, and so the perceptual noticing mechanism could not make use of the pattern, and in turn the use of the efficient strategy was not prompted. Alternatively, the pattern may only be useful for generating MIC if concepts central to DAC are known, without this the pattern may serve as a distraction. Similar explanations could be offered for the interaction between pattern and practice for the amount learned about the device.

In summary, the evidence reported here suggests that repeated practice of a known method can facilitate the generation of new strategies. A possible reason for this is that practice results in the elaboration of the user's device representation, which in turn supports strategy generation. These results challenge models of learning through practice that merely increase the efficiency of existing methods, and models of strategy change that fail to account for the role of practice.

Acknowledgements

Many thanks to Hans Neth and Gareth Miles for their useful comments on an earlier draft.

References

Anderson, J. R. (1995). *Learning and memory: An integrated approach*. New York: Wiley & sons.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*, 396-406.

Anzai, Y., & Simon, H. A. (1979). The theory of learning by doing. *Psychological Review, 86*, 124-140.

Bhavnani, S. K., & John, B. E. (1997). From sufficient to efficient usage: An analysis of strategic knowledge. *Proceedings of CHI '97*, 91-98.

Bhavnani, S. K., John, B. E., & Flemming, U. (1999). The strategic use of CAD: An empirically inspired, theory based course. *Proceedings of CHI '99*, 183-190.

Bhavnani, S. K. (in press). Strategic approach to computer literacy. *Proceedings of CHI 2000*.

Carroll, J. M., & Rosson, M. B. (1987). The paradox of the active user. In J. M. Carroll (Ed.), *Interfacing thought: Cognitive aspects of human-computer interaction*. Cambridge, M.A.: The MIT Press.

Compton, B. J., & Logan, G. D. (1991). The transition from algorithm to retrieval in memory based theories of automaticity. *Memory and Cognition, 19*, 151-158.

Crowley, K., Shrager, J., & Siegler, R. S. (1997). Strategy discovery as a competitive negotiation between metacognitive and associative mechanisms. *Developmental Review, 17*, 462-489.

Delaney, P. F., Reder, L. M., Staszewski, J. J., & Ritter, F. E. (1998). The strategy-specific nature of improvement: The power law applies by strategy within task. *Psychological Science, 9*, 1-7.

Logan, G. D. (1988). Towards an instance theory of automatization. *Psychological Review, 95*, 492-527.

Luchins, A. S. (1942). Mechanization in problem solving. *Psychological Monographs, 54*, 248.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the power law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition*. Hillsdale, New Jersey: Erlbaum.

Nilsen, E., Jong, H., Olson, J. S., & Polson, P. G. (1992). Method engineering: From data to practice. *Proceedings of ACM INTERCHI'93 Conference on Human Factors in Computing Systems*, 313-320.

Nilsen, E., Jong, H., Olson, J. S., Biolsi, K., Rueter, H., & Mutter, S. (1993). The growth of software skill: A longitudinal look at learning and performance. *Proceedings of ACM CHI'92 Conference on Human Factors in Computing Systems*, 149-156.

Olson, J. R., & Olson, G. M. (1990). The growth of cognitive modelling in human-computer interaction since GOMS. *Human Computer Interaction, 5*, 221-265.

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General, 126*(3), 288-311.

Rosenbloom, P. S., & Newell, A. (1986). The chunking of goal hierarchies: A generalized model of practice. In R. S. Michalski, J. G. Carbonell, & J. M. Mitchell (Eds.), *Machine learning 2: An artificial intelligence approach* . Los Altos, CA: Morgan Kaufmann.

Shrager, J., & Siegler, R. S. (1998). SCADS: A model of children's strategy choices and strategy discoveries. *Psychological Science, 9*(5), 405-410.

Young, R. M., & MacLean, A. (1988). Choosing between methods: analyzing the user's decision space in terms of schemas and linear models. *Proceedings of CHI '88*, 139-143.

A Potential Limitation of Embedded-Teaching for Formal Learning

Mei Chen (MeiChen@Vax2.Concordia.Ca)
Department of Education, Concordia University
1455 Maisonneuve Boulevard W., Montreal, Canada, H3G 1M8

Abstract

This paper presents a study that investigated the effects of two forms of "embedded teaching" on students' formal learning of high-school-level algebra. The term "embedded teaching" here refers to the presentation modes in which algebraic concepts, procedures, strategies, and principles are taught within the context of solving specific problems. Two forms of "embedded teaching" (i.e., program control, and learner-control) were compared to a conventional presentation mode in which various forms of algebraic representations were taught as a coherent system before students were introduced to problems. The three instructional modes were implemented in three versions of a computer algebra tutor. Three groups of high-school students (N=27) were randomly assigned to one of the three experimental conditions. Pre- and posttests were administered to measure changes in students' ability to construct different forms of algebraic representations, and their ability to make estimates using these different representations. A multivariate analysis of the pre- and posttest results indicates that overall student performance in all three conditions improved significantly on the two measures used (F (2, 23) = 46.6, p < 0.01). However, students in the conventional teaching condition achieved higher posttest scores (group mean = 86, and 93 on the measures of students' ability to construct algebraic representations and their ability to make estimates, respectively) than students in the two embedded-teaching conditions did (group mean for the two measures = 65 and 57 for the program-controlled condition; and group mean = 67, 59 for learner-controlled condition, respectively). Furthermore, all students in the conventional teaching condition completed the posttest successfully, compared to only 56% of students did so in each of the two embedded teaching conditions. Despite an overall disadvantage, 3 out of 9 students in each of the two embedded-teaching conditions received perfect scores whereas only 1 student in the conventional teaching condition did so. It seems that the conventional presentation mode provided an instructional context that enabled the majority of students to succeed whereas the embedded-teaching modes offered the conditions for certain students to reach their greatest potential but left others far behind.

This paper presents a study that investigated the effects of two presentation modes of "embedded teaching" on students' formal learning of high-school-level algebra. The term "embedded teaching" can be defined broadly or narrowly, depending on the purpose of the study. In the broad sense, most situated approaches that emphasize the learning of domain knowledge through expert-like activities and authentic problem solving in rich social, cultural and functional contexts can be thought as "embedded teaching." For example, the cognitive apprenticeship model (e.g., Collins, Brown, & Norman, 1989), problem-based learning (e.g., Barrows, & Tamblyn, 1980), goal-based scenarios (Collins, 1994; Schank, 1995), and case-based reasoning (Schank, 1995; Kolodner, 1994) can be all considered as "embedded teaching" (Hmelo, & Narayanan 1995). Although they can be characterized in many different ways and a wide array of features can be identified, situated approaches typically (a) focus on ill-defined, authentic problem solving tasks; (b) embed the learning of domain knowledge within a rich social and cultural context; (c) organize learning content and activities around problems or cases. Consequently, concepts, procedures, strategies, and principles are taught within the context of solving specific problems, as opposed to being taught as a coherent symbol system independently. Studies of "embedded teaching" in the past have often focused on the central role that social and cultural context played in learning, thus reflecting research traditions in anthropology and sociology.

The examination of social context is necessary for forming theories and good practice of situated learning. However, the focus of the study presented in this paper was placed on investigating the effects of "embedded teaching" from an instructional design perspective. That is, "embedded teaching" is defined as a narrow term that signifies the instructional techniques by which concepts, procedures, strategies, and principles are taught within a specific problem-solving context. The "embedded teaching" examined in this study represents merely an instructional feature but not the social and cultural characteristics of the situated learning approaches described above. The exclusive focus on the instructional feature of the "embeddedness" of the situated approaches in a study is justified for the following two reasons. First, as human beings and life-long learners, we have to complete a large part of our learning under solitary conditions due to the many practical limitations of getting together and learning with other people. Therefore, it is important to investigate what characteristics of instructional applications can effectively support the development of the individual knowledge in situations where rich social interaction and resources are lacking. Secondly, there is a formal system of symbols, or a structured body of scientific knowledge in a formal knowledge domain such as mathematics, physics, or chemistry. Therefore,

it is necessary to find out whether teaching concepts, procedures, strategies, and principles within the context of solving specific problems is more effective than teaching such knowledge as a coherent, representational system before introducing students to problems, as teachers often do in classroom teaching. The answer to this question is needed in order to design instructional applications, especially computer tutoring systems that effective foster individual learning in formal knowledge domains.

The Theoretical Background

A primary focus of analysis in cognitive science over the past three decades has been the processing structures of the brain and the symbolic representations of mind (Norman, 1993). Norman further characterizes such a research paradigm as studies of *symbolic representations* (Norman, 1993). Cognitive researchers have shown that structured, principle-based knowledge representation is a function of expertise (Chi, Feltovich, & Glaser, 1981). Therefore, one important goal of instruction is to help students form coherent and structured knowledge representations of the domain being studied (Glaser, 1989). Although they specify neither the elements of which instruction should be composed, nor the sequence in which the different instructional elements should be arranged, some cognitive theories seem to support an instructional approach by which the knowledge elements are structured and coherent, preferably proceeding from a declarative stage to a procedural stage. Kintsch's Construction-Integration (CI) model of discourse comprehension (Kintsch, 1988, 1998) and Anderson's ACT-R theory of cognitive skills represent two central tenet of this research paradigm.

It is apparent that Kintsch's CI model is a cognitive model of discourse comprehension rather than an instructional model, however, this model explains the ways that the mental representations are constructed when learning from text. According to Kintsch, comprehension (i.e., an important form of learning) is an interactive process between the learning materials and the mental models that learners formed on the basis of their prior knowledge. Kintsch suggests that the process of building a coherent mental representation from text depends on the learner's ability to recognize the structure of the text (Kintsch, 1998). Many studies have shown that the structural features of a text (e.g. coherence, the use of the outlines, headings) can help learners identify the structure of the text thus having a significant impact on learners' memory, comprehension, problem solving, and transfer (Kintsch, 1997, in Kintsch 1998; Mayer, 1989; 1997). It is postulated that the structural features of a text may enable learners to form a coherent knowledge representation of the material to be studied, the coherent representation formed, in turn, can guide the processes of selecting, interpreting, organizing, and integrating the subsequent information to be studied (Mayer, 1989). As a result, the structure feature of text can has impact on learning, especially for learners who lack prior knowledge of the domain (Kintsch, 1997, in Kintsch 1998).

Furthermore, Anderson's ACT-R theory asserts that the acquisition of cognitive skills proceeds from a declarative stage to a procedural stage (Anderson, 1983, 1993). Although both Kintsch and Anderson state that learning should be embedded in the context of meaningful activities, their theories seem to support a pedagogical approach in which learning proceeds from instruction that focus on the structured knowledge representations, to problem-solving activities that emphasize the use of the knowledge learned — a method that is frequently employed by teachers in their classroom teaching.

In contrast to the emphasis on the structured knowledge representations and rigid instructional sequence, situated learning approaches give great importance to the functional use of knowledge thereby the learning materials and activities are organized around problems or cases. The theoretical assumption underlying such approaches can be refered to as situated theory. From the situated perspective, thinking, knowing, and learning are situated within a particular context of intentions, social partners, and tools (Resnick, Levine, & Teasley, 1991; Greeno, 1997). Therefore, internal cognitive activities such as perceiving, understanding, remembering and reasoning are shaped and given significance within the context of activities (Greeno, Collins, & Resnick, 1996). The situative view challenges standard pedagogical practice for paying too little attention to the processes employed by experts to solve complex, realistic problems (Collins et al., 1989; Resnick et al., 1991). Situative theorists criticize learning opportunities provided to students within the scope of typical school activities as mostly involving memorization of factual knowledge and the rote manipulation of symbols and equations. As a result, the kind of knowledge that students acquire in conventional teaching often remains "inert" and can't be applied to other relevant problem-solving situations (Collins et al., 1989; Resnick et al., 1991). It is asserted that "embedded teaching" can facilitate students' development of problem-solving skills and reasoning strategies more effectively. It can also enhance students' conceptual understanding because knowledge is immediately used within a relevant context (Collins et al., 1989). A variety of pedagogical approaches that have been developed emphasize such situative nature of learning and cognition (e.g., Collins et al., 1989; Collins, 1994; CTGV, 1993; Schank, 1995; Kolodner, 1994).

An interesting phenomenon occurred in the debate over symbolic representations versus situated action is that, as Norman points out, the cognitive and situative researchers find different sets of observations to be

interesting and important (Norman, 1993). The same statement may also apply to the argument over conventional school teaching versus situated approaches to learning. Conventional teaching normally focuses on the understanding of symbolic representations and the command of symbolic manipulations whereas situated approaches typically places their focus on the development of students' ability to formulate and solve ill-structured, authentic problems. It is important to keep in mind that, in a knowledge domain such as mathematics, physics, or chemistry, there is a formal system of symbols, or a structured body of scientific knowledge. Therefore, an expert model of knowledge in such domains may consist not only of the strategies that are used in expert-like performance, but also of structured, principle-based knowledge representations of symbolic systems. Correspondingly, students should be encouraged and required to develop the full range of knowledge and skills in the domain, including the ability to understand the symbol systems correctly and manipulate symbols intelligently, the ability to communicate ideas scientifically, the ability to formulate and solve ill-structured problems proficiently and, ultimately, the ability to participate in expert practice in the real world. Therefore, an important question concerning "embedded teaching" is to understand whether it is indeed a more effective technique than the conventional method for teaching formal knowledge. The study presented in the paper attempts to find out whether the embedded teaching as a presentation mode is more effective than conventional teaching for computer-assisted learning of high-school-level algebra.

Methods

This study compared the effects of two presentation modes of "embedded teaching" (i.e., program control, and learner-control) to that of a conventional presentation mode. The three experimental modes were implemented in three versions of a computer tutor that was designed to teach linear functions to grade ninth students. Several features of the computer tutor are important for interpreting the results of the study. First, a cognitive task analysis was conducted to identify the specific elements of knowledge that students need in order to (a) construct multiple forms of algebraic representations (i.e., tables of values, graphs, and equations), and (b) make estimates using such representations (e.g., finding the price of ordering a given number of music CDs based on the relations expressed in a graph of linear functions). Second, the computer tutor employed a basic instructional model that consisted of instruction, demonstration, and practice. In addition, everyday-life scenarios were incorporated into the instruction, demonstration, and practice whenever possible, to help students make connections between their knowledge of the everyday life and the formal algebraic representational system that they were to learn. Furthermore, various media formats (text, graphics, and animations) were combined to describe some complex concepts, principles, and procedures. Finally, all three versions of the computer tutor utilized the same instructional materials, practice exercises, and media formats, varying only in terms of the "embeddedness" and "learner control". The following is a brief description of these conditions.

The first version of the tutor was designed as a base line for making the comparison (Condition 1). In this version, the computer tutor first presented instruction about the different forms of algebraic representations, explaining what are tables of values, graphs, or equations, and the relationship between the different forms of representations. After the instruction, the computer tutor then provided examples illustrating how to construct each of the different forms of algebraic representations and how to convert these representations from one form to another.

Two forms of "embedded teaching" were examined: a program-control instruction (Condition 2), and learner-control instruction (Condition 3). Both conditions presented the instruction about the different forms of algebraic representations (i.e., what are tables of values, graphs, or equations) within the context of constructing these representations. Because it was the instructional designer of the computer tutor who decided the occasions to introduce the relevant algebraic concepts and principles, students had to follow a fixed sequence imposed by the computer tutor in the program-controlled instruction (Condition 2). The ways that the computer tutor presented various types of algebraic knowledge might not meet the needs of individual students. Therefore, a learner-controlled hypermedia environment (Condition 3) was developed to enable students themselves to determine what, and when to consult a particular type of knowledge. In the learner-controlled hypermedia environment, students engaged directly in problem-solving activities without receiving any prior instruction or demonstration, but they could receive relevant instructions and demonstrations using hyperlinks. It is necessary to indicate that for both Condition 2 and 3, the instruction on algebraic representations was broken into small independent chunks to enable the computer tutor or students to access the relevant knowledge elements in the context of solving specific problems. This might create stronger links between particular knowledge chunks and their applications, but weaker links between the different forms of algebraic representations.

All three conditions included a practice session in which two problems similar to the one used in the demonstration were presented. Students could check their answers, or get the correct answers if them failed to provide the correct answers.

Participants

A public English school in suburban Montreal was

selected as the setting of the experiment. An effort was made to recruit as many participants as possible. As a result, two types of students constituted the sample pool for the experiment: (a) ninth graders in a below-average class, who had just finished learning linear functions one week prior to the pretest, but who had difficulties in math classes; and (b) eighth-graders from two regular math classes who had some knowledge of algebra, but hadn't explicitly learned about linear functions. A pretest was administered to assess the prior knowledge that students had on linear functions. Twenty-seven students who scored under 60 on a scale of 100 were selected to participate in the experiment.

Procedures

The twenty-seven students were divided into three groups based upon their pre-test scores, the grades and gender were balanced in each group. Each group was then randomly assigned to one of the three experimental conditions. Some adjustments were made to accommodate the schedules of the participants. The experiment consisted of two 45-minute learning sessions over a two-week period in the school computer lab. The algebra tutoring program was used as the sole source of instruction and each student learned alone with one version of the computer program. Instruction on the use of a particular version of computer program was presented by the computer at the beginning of the learning sessions. In order to understand types of learning activities in which students engaged, student-computer interaction was recorded during the experiment. To conclude the experiment, a posttest was administered one week after the last learning session. The posttest consisted of (a) two word problems to assess students' algebraic skills and the transfer of such skills, and (b) questions designed to assess students' understanding of algebraic concepts and rules. After students had completed the posttest, they were asked to complete an attitude questionnaire concerning their learning experience with the computer tutor. In addition, three students who received the best posttest scores were interviewed one week after the posttest. They were asked first to orally answer some structured questions in attempt to understand their conceptual knowledge, problem-solving and reasoning strategies, then to translate the algebraic representations from one form to the other in "backward" fashion in order to determine whether student performance was based on the execution of a set of "rote" procedures or on the understanding of the algebraic relations.

Results

The independent variable of this study is the presentation modes: two forms of "embedded teaching" were compared to a conventional teaching mode. The dependent variables include the pretest and posttest measures that reflect the development of students' algebraic knowledge and skills: (a) the measure on the ability to construct different forms of algebraic representations; and (b) the measure on the ability to make estimates using the relationships expressed by such representations. In addition, several other dependent variables were measured in the posttest but not in the pretest: (c) the measure of the transfer of algebraic knowledge and skills to the solution of a problem in an unfamiliar situation, (d) the measure of understanding of the relationships between different forms of algebraic representations by asking students to construct the various forms of algebraic representations that they had learned in "backward" fashion, and (e) the assessment of students' conceptual understanding of the key algebra concepts and principles by asking them to fill in the answers to some conceptual questions. In addition, students offered their opinions about their learning experience by answering an attitude questionnaire.

Two types of "process" data were also collected to determine how the three different teaching modes affected students' learning of well-structured algebraic tasks: (a) the "dribble" file recordings of the history of student-computer interaction; and (b) the explanations that the selected students provided in response to structured interview questions.

The data-analysis strategy consists of two steps: The first step is comparing the learning outcome measures to determine whether the three different presentation modes have different effects on students' learning outcomes. A multivariate analysis (MANOVA) is employed to analyze the pre- and posttest scores on students' ability to construct different forms of algebraic representations and their ability to make estimates using these representations. This is followed by a qualitative analysis of the learning process measures to understand what may have contributed to the effects observed. The results of this study were reported in details elsewhere (Chen, 2000). This paper only highlights the findings concerning the following questions:

(1) did the three versions of computer tutor improve student performance on well-structured algebraic tasks?

The population mean scores on the measure of students' ability to construct algebraic representations was raised from 22.5 to 72 on a scale of 100. Similarly, the population mean score on the measure of students ability to make estimates using these representations was raised from 27.1 to 70. Table 1 presents the pre- and post-test scores of these two variables for all three conditions. A multivariate analysis (MANOVA) of the pre- and posttest scores indicates a significant overall effect on the two measures used ($F(2, 23) = 46.6$, $p < 0.01$). Therefore, it seems that students improved not only their ability to construct algebraic representations, but also their ability to make estimates using the different forms of algebraic representations. However, it

Table 1: The pre- and post-test scores of two measures of students' algebraic abilities (N=27).

Presentation Modes	Constructing Representa-tions		Making Estimates	
	Pre-test	Post-test	Pre-test	Post-test
C1: Conventional	17.9	86.4	25.9	92.6
C2: Program-control	29.8	64.8	33.3	57.4
C3: Learner-control	19.9	67.3	22.2	59.3
M	*22.5*	*72.8*	*27.1*	*70*
SD	*21.4*	*31.6*	*33.4*	*37.2*

should be noticed that the standard deviations are considerably large in both the pre- and post-test scores (SDs range from 21 to 37), reflecting big individual differences in the participants' initial algebraic abilities, as well as the improvement of such abilities.

(2) Did "embedded teaching" modes promote better learning outcomes than conventional teaching?

The MANOVA rejects the hypothesis of no difference between the conventional teaching mode and the program-controlled embedded-teaching mode on students ability to construct algebraic representations and their ability to make estimates using these representations ($F(2, 23) = 3.7$, $p < 0.05$). Similarly, the hypothesis of no difference between the conventional teaching and the learner-controlled embedded-teaching on the same two dependent variables is rejected ($F(2, 23) = 4.1$, $p < 0.05$). Therefore, students' learning outcomes in the embedded teaching modes differ significantly from that in the conventional teaching mode. However, contradicting the hypothesis that the embedded teaching is more effective than conventional teaching, this study reveals that students produced better learning outcomes in the conventional teaching mode. Students in conventional teaching mode also showed superiority in terms of their performance on transfer task, and measure on conceptual understanding. It is important to indicate that all students in the conventional teaching condition (including those who had difficulty in the regular math classroom) successfully passed the posttest, whereas only 56% of students passed the post-test in the other two conditions. However, three out of nine students in each of the embedded teaching modes received the perfect scores whereas only one out of nine students in the conventional teaching mode was able to do so. It seems that the conventional teaching provides an instructional context that enables the majority of students to succeed whereas the embedded-teaching conditions offer the conditions for certain students to reach their greatest potential.

(3) Did "learner-controlled" instruction enhance student learning in a "embedded-teaching" mode?

No difference is found between the learner-controlled and the program-controlled embedded-teaching modes on the measures of students' ability to construct algebraic representations and their ability to make accurate estimates ($F(2, 23) = 0.46$, $p > 0.05$). Thus, the level of student control (i.e., being able to determine whether or when to consult the different types of knowledge available) did not influence students' learning outcomes. However, the examination of student-computer interaction reveals what differentiates the high achievers from the low achievers is the navigation strategies that they employed when they were given control over the navigation paths. The high achievers tended to complete each topic that they reviewed whereas the low achievers tended to switch from topic to topic, indicating a lack of mindful engagement.

(4) What factors may have contributed to different learning outcomes observed?

The analysis of the interview protocol reveals that the strategies that students employed to solve problems are linked to the types of conceptual understanding that they developed. The interview also indicates that successful task performance is associated with conceptual knowledge that is retrievable and, more importantly, that is coupled with appropriate reasoning strategies.

The examination of student-computer interaction and students' self-reports indicates that students' posttest performance relates to neither the amount of time that students spent on tasks, nor to the amount of instructional materials that they reviewed. However, the navigation strategies that students used in the learner-controlled condition seem to relate to the outcomes of learning. The surfer's strategy (i.e., switching from topic to topic without completion) is directly linked to poor posttest performance. The "learning curve" derived from the dribble file recordings further indicates that (a) significant learning had took place through instruction and demonstration and, (b) a certain amount of learning took place through practice, however, the role of practice was not as predominant as what is usually believed. The questionnaire and interview data indicate that student attitudes toward their learning with the computer tutor program are very positive, irrespective of the teaching conditions. Therefore, the superior posttest performance of the students who learned under the conventional teaching condition seem primarily due to the characteristics of this presentation mode rather than any other factor.

Discussion

This section will discuss briefly two important findings of this study. First, this study shows that students' algebraic knowledge and skills have improved significantly over time. Such a finding suggests that

computer-based learning environments that employ an adequate instructional model, incorporate authentic problem scenarios, provide rich learning activities, and use multimedia to illustrate abstract concepts and procedures, can effectively enhance student learning of algebra. Second, this study indicates that students in the conventional teaching condition generally learned better than students in both program-controlled and learner-controlled embedded teaching conditions did. It is speculated that the conventional teaching condition may enable the majority of students to develop a coherent mental model of the algebraic representations. The coherent knowledge representations that students developed, in turn, may help them better interpret the goals and functions of the subsequent procedures and effectively direct their attention to the key task elements in the step-by-step demonstration performed by the tutor, as well as in their own problem-solving exercises. Furthermore, this study shows that significant learning had taken place through instruction and demonstration prior to the practice of the exercises. The findings of this study seem to support the notion that learning is a sense making process that involves construction and integration of mental representation of the materials being studied, as explained by Kintsch's CI model (Kintsch, 1989, 1998). The findings of this study suggest that instructional designers need to consider the coherence of the symbolic representations when applying "embedded teaching" and "learning by doing" principles to design instruction applications for formal learning. This study also showed that the percentage of students who received perfect posttest scores is higher in the embedded teaching conditions than that in the conventional teaching condition. It seems that the problem-solving context and the higher level of learner control enables some students to reach their greatest potential. Therefore, problem-solving context may indeed more effectively facilitate the development of conceptual understanding, problem-solving skills, and reasoning strategies for some students. In addition, some students are able to focus on their own weaknesses when they have control over what to learn.

Some limitations of this study include the use of a small sample of the participants and a lack of aptitude and learning strategy measures prior to the experiment. A further direction of this research is to incorporate other features of successful "embedded teaching" approaches (e.g., using diversified and contrasting cases to enable students to generate the "rules of thumb") into the computer algebra tutors described here to investigate how these features can support both collaborative inquires and individual learning. The goal is to understand the pedagogical requirements for implementing particular instructional approaches and to explore their limitations and strengths for teaching in a given domain. The findings of such studies will certainly have important implications for the development of instructional theories and applications. However, precautions are in order when discussing the pedagogical approaches taken by teachers in classroom situations. This is because good teachers adjust their teaching techniques according to their assessment of students and their monitoring of the on-going instructional processes—they rarely use an instructional technique exclusively even though they may firmly believe in a particular pedagogical approach .

Acknowledgements

The research was supported in part by Fonds FCAR. Special thanks to my husband Michel Décary for his contributions and support. The author would also like to thank Dr. Frederiksen, Dr. Breuleux and Dr. Lajoie for their thoughtful comments and suggestions.

References

Chen, M. (2000). *The impact of three instructional modes of computer tutoring on student learning in algebra*. Unpublished doctoral Dissertation, McGill University, Montreal.

Chi, M. T. H., Feltovich, P., & Glaser R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121-152.

Collins, A., Brown, J. S. and Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.) *Knowing, learning and instructions: Essays in honor of Robert Glaser*. Hillsdale, NJ: Erlbaum.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. In D. C. Berliner & R. C. Calfee (Eds.) *Handbook of educational psychology* (pp. 15-46). London, UK: Prentice Hall.

Hmelo, C. E., & Narayanan, N. H. (1995). Anchors, cases, problems, and scenarios as contexts for learning. *Proceedings of the 17th Annual Meeting of the Cognitive Science Society* (pp. 5-8).

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*, 163-182.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press, Cambridge.

Kolodner, J. L. (1994). *Case-based reasoning*. San Mateo, CA: Morgan-Kaufmann.

Mayer, R. (1989). Models for understanding. *Review of Educational Research, 59(*1), 43-64.

Norman, D. A. (1993). Cognition in the head and in the world: An introduction to the special issue on situated action. *Cognitive Science, 17*(1), 1-6.

Resnick, L. B., Levine, J. M., & Teasley, S. D. (1991). *Perspectives on socially shared cognition.* Washington, DC: American Psychology Association.

Schank, R. C. (1995). *Engines for Education*. Hillsdale, NJ: Erlbaum.

Drawing out the Temporal Signature of Induced Perceptual Chunks

Peter C-H. Cheng (peter.cheng@nottingham.ac.uk)
Jeanette McFadzean (jmc@psychology.nottingham.ac.uk)
Lucy Copeland (lap@psychology.nottingham.ac.uk)
ESRC Centre for Research in Development, Instruction and Training, Department of Psychology, University of Nottingham, Nottingham, NG7 2RD, U.K.

Abstract

The effect of chunking in the process of drawing was investigated using a task domain consisting of simple hierarchically organized geometrical patterns, which participants learnt to draw. The study focused upon the latencies between drawing actions. A new technique for the identification of chunks was devised, based on patterns in the magnitudes of latency. The technique was significantly better than the use of a fixed latency threshold. It was discovered that there was a strong temporal signature of the underlying chunk structure and that effects of learning were evident.

Introduction

The concepts of chunking and the limited size of memory span, first proposed by Miller (1956), underlie many modern theories of human cognition. The phenomenon has been verified in many domains (Vicente, 1988), and at most levels of cognitive processing in both humans and nonverbal organisms (Terrance, 1991). Given the pivotal role of chunking it is, perhaps, surprising that there has been little research on the role of chunking in drawing. There has been some research on: low level motor behaviour constraints on drawing (Van Sommers, 1984), the functions of drawing in high level cognitive tasks such as design (Akin, 1986) and, drawing as a reflection of children's cognitive development (Goodnow & Levine, 1973). However, direct investigations of the role of chunking in the process of drawing are absent.

We are conducting studies that begin to address this deficiency in the understanding of this prominent human ability. Our approach is to have participants learn specially designed named geometric shapes, from verbal labels, which they then reproduced from memory – drawing out induced perceptual chunks. This paper focuses on whether chunks are apparent in temporal characteristics of drawing. Specifically, we have discovered that the absolute duration of pauses between drawing actions, the latencies, reflects the hierarchical structure of induced chunks and reveals the effect of learning by the composition of chunks. Further, we have found that local maxima in the latencies are better discriminators of boundaries between separate chunks than a fixed latency threshold.

Previous work on chunking and drawing will first be discussed to set the context for this work.

Chunking and Pausing Behaviour

The idea that latencies or pauses might be a means by which one can segment data in order to discriminate chunk boundaries, arises from research conducted by Chase and Simon (1973). They defined an operational method by which to characterize chunks. In recall and memory tasks the latency distributions for between-glance placements of chess pieces, which were taken to indicate boundaries between chunks, were significantly longer than within-glance placements, which were taken to indicate items within a chunk. Hence, items with pauses below a certain threshold could be considered as within a chunk and items with longer pauses above the threshold could be considered to be between chunks. The use of thresholds as one means to distinguish chunks has been supported by studies in domains such as: Chess (Chase & Simon, 1973; Gobet, 1998), Go (Rittman, 1976,1980), and electronic circuits (Egan & Schwartz 1979).

A significant pause can be defined as a latency greater than a static threshold typically within the range 2 to 5 seconds (Card, Moran, & Newell, 1983). Although, in studies of drawing, researchers have used pauses as low as 1 second to segment data into chunks (Akin, 1986; Ullman, 1990).

However, there are difficulties with the use of latency thresholds to differentiate chunk boundaries (Holden, 1992; Gobet, 1998). Firstly, there is no one threshold that holds across different task domains (Chase & Simon 1973; Egan & Schwartz, 1979; Akin, 1986; Ullman, 1990; Gobet & Simon, 1998). However, a threshold can be found by training participants in a domain and then testing them (Reitman, 1976). Secondly, it has been observed that when learning takes place, as in the transition from novice to expert, latency times for chunk boundaries decrease (Chase & Simon 1973; Reitman, 1976; Egan and Schwartz, 1979). Thresholds must be changed dynamically over time to cope with individual differences. Thirdly, for memories that are organized hierarchically (Palmer, 1977), the higher the chunk is in the hierarchy the more subchunks it contains and the longer it takes to recall (Reitman, 1976). A single threshold might elicit chunks at one level but not its subchunks or higher order chunks.

This paper proposes an alternative approach to thresholds for the identification of chunks using

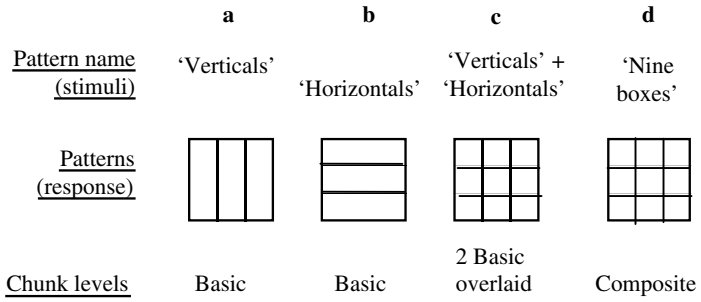


Figure 1. Examples of types of patterns from the shape drawing domain.

latencies between drawing actions. By focusing on patterns over successive latencies the new technique can overcome some of the limitations of fixed thresholds.

Here, we define latency for a particular element as the time between lifting the pen off the paper at the end of one element and the time at which the pen touches the paper again at the beginning of marking the current element. The same holds for mouse button up and down actions when dragging a line on a computer screen.

The Nature of Drawing

Intuitively and theoretically there are various reasons to believe that understanding the role of chunks in the process of drawing will be a challenge. First, consider the recall and the drawing of a perceptual chunk given a verbal label for that chunk. A succession of processes are involved, including: the recall of the chunk, the planning of the order in which to produce the elements of a chunk, the planning of where to draw each individual element, and, the execution of the motor actions to make a mark for the element. It seems likely that such a sequence of processes would hide any hierarchical organization of chunks in long-term memory. Second, it appears unlikely that these processes would occur in a strictly serial manner and they are likely to be interleaved to different extents. This will probably mask any attempt to analyze the underlying structure of chunks. Third, the process of planning might in itself interfere with the recall of chunks and so potentially prevent each chunk from being recalled in a single burst of activity (Reitman, 1976). One might reasonably assume that analysis of latencies within this area would reflect planning and action, rather than chunking. Fourth, the processes of mark making, including subjects sensitivity to methods of motor efficiencies (Akin, 1986), might interact with the recall of chunks. For example, the speed of making a mark may vary with the hierarchical chunk level of the current element being drawn and so interfere with the apparent recall latency of the next element.

Despite all these reasons, the experiment reported here demonstrates that the duration of pauses between the drawing of individual elements is highly indicative of the structure of chunks in memory. It appears that far from diluting any information about the underlying organization of chunks, the duration of latencies in the process of drawing seems to provide a temporal signature for perceptual chunks.

The next section presents the drawing domain and task used in the experiments. The following section describes the discovery of patterns in the latencies that were highly suggestive of a temporal signature of chunks. The experiment and results that demonstrate the reality of these patterns are then considered in turn. The implications of the findings are considered in the final discussion section.

Domain, Stimuli and Tasks

To study the behavioural manifestations of chunking in drawing a special 'shape' domain consisting of named geometric patterns was devised; examples are shown in Figure 1. Initially participants were taught six basic patterns, such as Figure 1a and 1b, and they drew several examples of them when given their names. They were then shown pairs of names of basic patterns to draw overlaid upon each other; Figure 1c. These composite patterns were then separately named; Figure 1d. The composite patterns consist of four lines and a typical drawing task involves drawing a sequence of different composite patterns in a row beside a written list of names.

Features of the design of the domain that make it highly suited to the study of chunking behaviour are: (1) the use of simple predefined shapes to make errors in recall or drawing easily identifiable, (2) the definition of a fixed hierarchy of patterns, with known nesting of levels and no overlapping of elements over chunks at the same level, (3) the participants learn the patterns, so the specific chunks and their organisation is known *a priori,* (4) verbal labels are used as stimuli to make participants recall the perceptual chunk from long-term memory, (5) the composite chunks consist of a small number of sub-chunks to keep demands on working memory low, (6) the outline square is drawn before each pattern to ensure drawing processes are fully engaged when the pattern is produced.

The domain has three chunk levels: (level 1) lines within chunks, (level 2) basic patterns, (level 3) composite patterns. Thus, every line was coded depending on the order in which they were drawn. The first and second drawn lines of a two line basic pattern were coded level 2 and 1, respectively. The code for the four lines of two basic patterns overlaid was 2-1-2-1, respectively. The four lines of a composite pattern were coded 3-1-2-1, respectively, on the assumption that the composite consisted of two sub-chunks.

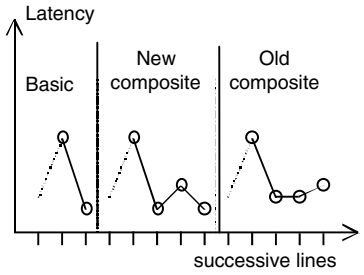


Figure 2. Temporal signature for different chunks

Motivating Observations and Hypotheses

The experiment reported below consists of two experiments (taken together here for the sake of brevity and coherence). The first was a pilot in which latency and other measures were examined in an exploratory manner. In graphs for various measures based on data from each individual participant on a single task (i.e., raw un-aggregated data), it was noticed that certain patterns of latencies appeared to be common and were related to the participants' induced chunks. Figure 2 illustrates the patterns found. Local maxima in latencies, *peaks*, tended to be associated with the first line of basic and composite chunks. A peak was operationally defined as any latency whose magnitude was at least 10% greater than the mean of the preceding and following latencies. With new just learnt composite chunks there were two peaks, matching the two sub-chunks, with the second peak being smaller. With old composites that had been drawn many times, and so learnt well, the second peak tended to disappear.

Although the patterns illustrated in Figure 2 were not universal they were sufficiently frequent to suggest that some temporal signature of chunks would be found by analyzing latencies. In particular, we propose three hypotheses: (H1) Peaks may be an effective way to discriminate chunks. Are peaks better than fixed latency thresholds for identifying chunk boundaries? (H2) Hierarchical chunk levels may be reflected in the absolute magnitude of latencies. (H3) The learning of chunks may be apparent in changes of latencies over time. Further, (H4) if the temporal signature of perceptual chunks is real then it should be apparent when different drawing media are used. The purpose of the experiment was to test these hypotheses.

Experiment

The first two hypotheses were tested by using the shapes domain described above. The third hypothesis concerning learning was tested by comparing performance over two successive sessions in which the same patterns were learnt and reinforced. The fourth hypothesis was tested by using two different drawing interfaces – pen and paper drawing versus keyboard and mouse driven on-screen computer drawing, henceforth *freehand* and *computer* groups, respectively.

The participants were unpaid volunteers, 4 male and 4 female aged 30-45. Equal numbers were assigned to the computer and freehand groups.

Apparatus

The computer drawing used a specially written program on a Macintosh G3 computer. To draw a line, participants first used the keyboard to select the type of line to be drawn (i.e. horizontal, vertical or diagonal) by pressing a key. The line was then drawn using a standard mouse dragging action, with the line "rubber banding" between the endpoints.

The freehand drawing used a high spatial and temporal resolution drawing tablet (Wacom UD tablet) connected to a PC computer running a specially written data capture and analysis program.

In both cases the computers recorded detailed spatial and temporal data to enable the drawn patterns to be identified and for the latencies to be found.

Procedure

Participants were tested individually and each completed two sessions. The participants were given a period of familiarization with the given drawing apparatus. In session 1, in order to learn the patterns participants completed drawings of several basic patterns. This was followed by a further 6 drawings of both and basic and composite patterns. In the session 2, there were 18 drawings consisting of multiple patterns. The stimuli were presented on printed sheets or by verbal instructions.

Results

Peaks versus thresholds: H1 and H4

Consider first the overall distributions of latencies for elements within chunks (level 1) and between chunks (levels 2 and 3). For data aggregated over participants in the same group and over all the tasks in each session, Table 1 shows various measures for these distributions. Between chunk latencies are greater than within chunk latencies, across session and interface type. All the distributions are skewed towards lower latency values. This pattern is similar to that found in other domains (e.g., Chase and Simon, 1973; Reitman, 1976; Egan and Schwartz, 1979), so it is appropriate to analyze this

domain using thresholds to identify chunk boundaries. As latency distributions were skewed, median latencies rather than the mean latencies were used in the analysis.

As expected, the median latencies were shorter for freehand drawing versus computer drawing because of the extra decisions and motor actions required with the computer drawing interface.

Table 1: Between and Within Chunk Latency Distributions (milliseconds)

	Group	Computer		Freehand	
	Session	1	2	1	2
Between Chunk	N	217	309	197	315
	Mean	1647	1188	2017	899
	Median	1347	931	989	620
	SD	1237	892	2985	1677
Within Chunk	N	214	325	186	286
	Mean	814	686	1113	413
	Median	681	665	475	389
	SD	958	340	3657	169

As the number of chunks is defined by the stimuli set an 'optimum threshold' can be set to distinguish chunk boundaries on an individual participant and session basis. The threshold is set so that the number of items above the threshold equals the number of known chunk items. Table 2 gives the thresholds found for each participant. Note the differences across sessions and the differences between individuals within sessions. As would be expected, the threshold for free hand is generally less than that for computer drawing.

Table 2: Optimal thresholds (milliseconds) for each participant

Computer sessions			Freehand session		
	1	2		1	2
P1	600	800	P5	400	400
P2	1400	800	P6	600	600
P3	1400	1200	P7	800	800
P4	1600	1200	P8	600	600

How do peaks compare with the use of latency thresholds as methods to discriminate chunks? Information theory (Wickens, 1993) provides a convenient way to measure how well each method performs by treating each as a system that is attempting to transmit information about items within chunks and items at the boundaries between chunks. By using conditional probabilities, information theory takes into account not only true positives and negatives (e.g., peak–>between chunks, ~peak–> within chunk) but also false positives and negatives (peak–>within chunk, ~peak–>between chunks). Using the optimal thresholds given above and the prior knowledge about which items were chunks or not, the correctness of each individual identification was determined. The same was done for peaks. Hence, the 'quality' (information transmission/channel capacity) of each method was computed (Wickens, 1993). The ideal quality, for all participants across the sessions, was almost unity because the numbers of within and between chunk items were nearly equal.

Table 3: Quality of chunk discrimination by the two methods (bits)

	Peaks		Threshold	
	Session 1	Session 2	Session 1	Session 2
Computer				
P1	0.628	0.444	0.398	0.247
P2	0.503	0.186	0.155	0.080
P3	0.306	0.195	0.261	0.092
P3	0.355	0.117	0.247	0.071
Freehand				
P5	0.460	0.595	0.133	0.274
P6	0.400	0.620	0.168	0.331
P7	0.671	0.622	0.321	0.205
P8	0.410	0.238	0.217	0.138

Table 3 shows that for each participant in each session under each drawing interface, the quality of the discrimination with peaks was better than that using fixed latency thresholds. Although there were just four participants in each group, two-tail t tests were performed to determine whether the peak method gave a significantly higher quality of discrimination than the threshold method. This was indeed the case for the freehand drawing in sessions 1 and 2 ($p=.005$ and $p=.02$, respectively) and computer drawing in session 2 ($p=.018$). For computer drawing in session 1 the difference was approaching significance ($p=.07$).

The results demonstrate: (H1) that peaks are a more effective way to discriminate chunk boundaries, and (H4) the temporal signature of perceptual chunks is apparent across different drawing media.

Chunking levels and learning: H2 & H3

Figures 3 and 4 show graphs of the median latencies against different chunk levels for each drawing interface and across each session. With one exception, for every participant in each session with both drawing interfaces, the median latencies increased with increasing chunk level. Using Page's test for ordered median alternatives as applied to the different levels of the chunk hierarchy (levels 1, 2, and 3) there was a significant increasing trend in the latencies for the computer drawing in both sessions; in both cases $L=56$, $p=.001$ ($n=4$, $k=3$). Similarly for Freehand drawing in session 1 $L=56$, $p=.001$ and session 2 $L=55$, $p=.01$ ($n=4$, $k=3$). The difference between the medians holds not only at the group levels but also at an individual level. Using the data for each participant, the Kruskal-

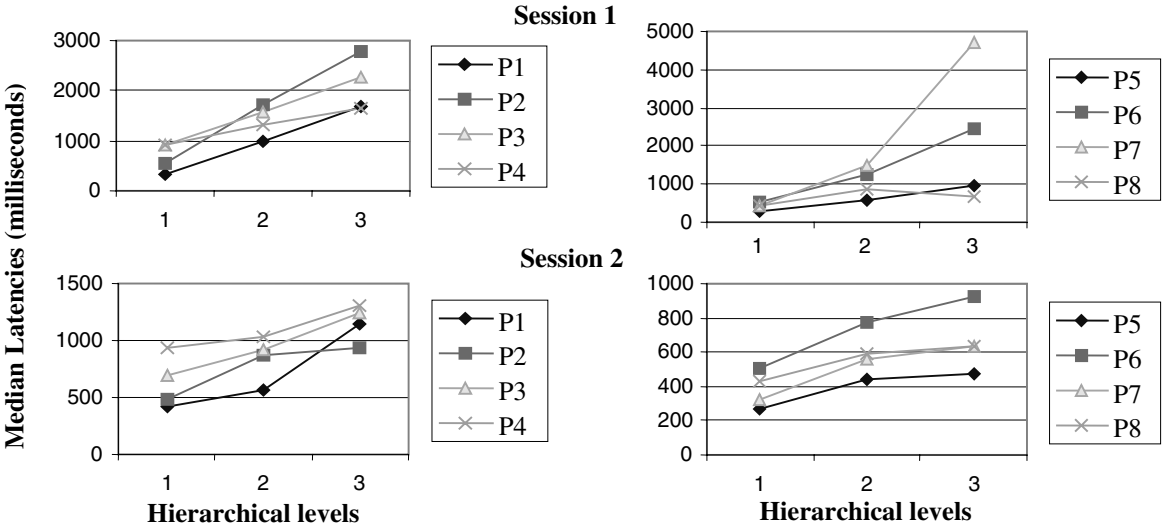


Figure 3: Computer drawing

Figure 4: Freehand drawing

Wallis-H test was used to test whether the latency distributions for the hierarchical levels were significantly different. As shown in Table 4, the results of the test for all participants were significant in both sessions and regardless of the mode of drawing. Comparing the graphs in Figures 3 and 4 across the two sessions for each mode of drawing, it is clear that the magnitudes of latencies drop. (Note that the latency scale ranges differ.)

The results demonstrate: (H2) that the magnitude of latencies reflect the hierarchical chunk level.

Table 4: Analysis of participants' latency distributions over the hierarchical levels; Kruskal-Wallis H

Mode of Drawing		Session 1		Session 2	
	Participant	n	$\chi 2$	n	$\chi 2$
Computer	P1	118	60.3^{+}	184	61.48^{+}
	P2	102	20.5^{+}	158	21.0^{+}
	P3	105	28.7^{+}	145	24.5^{+}
	P4	106	19.9^{+}	146	15.2^{+}
Freehand	P5	92	21.3^{+}	124	45.3^{+}
	P6	107	24.0^{+}	187	88.9^{+}
	P7	91	32.3^{+}	175	59.4^{+}
	P7	93	29.1^{+}	115	29.9^{+}

$^{+}p \leq .001$, df=2 in all cases

Table 5 presents median latencies for participants using each mode of drawing for each chunk hierarchy level and summarises the analysis. The latencies decreased over sessions regardless of the hierarchical level. The differences between participants performance over the two sessions was assessed by applying the Mann-Whitney U test (one-tailed); for the freehand drawing group the decrease in median latencies is significant at all chunk levels and for the computer drawing group the decrease is significant at chunk level-2 and level-3.

The results demonstrate: (H3) that the learning of chunks is apparent in the changes of latencies over time.

Table 5: Comparison of the latencies between sessions at each hierarchical level

Mode of Drawing & measure	Hierarchical levels					
	1		2		3	
	S1	S2	S1	S2	S1	S2
Computer						
Median	681	665	1131	865	1720	1114
N	542		401		117	
U	33523		14540^{+}		1079^{+}	
Z	-0.88		-4.44		-3.09	
Free-hand						
Median	472	389	989	584	1042	658
N	473		341		110	
U	19664^{+}		12228^{+}		626.5*	
Z	-4.869		-6.591		-1.88	

*p<.05, $^{+}p \leq .001$

Discussion

A specially designed geometric shapes domain has been used to study chunking behaviour in drawing. Participants learnt named patterns that were assumed, reasonably, to have been stored in memory as induced perceptual chunks. The differences in the distributions of recall latencies for elements within chunks and those between chunks is similar to patterns of latency distributions found in other domains (e.g., Chase and Simon, 1973; Reitman, 1976; Frey, 1976; Egan and

Schwartz, 1979; Akin 1986; Ullman, 1990; Gobet, 1998). Similarly, optimal latency thresholds that could be used to identify chunks were found to vary with participants, depending on the nature of the drawing interface and on the effect of learning.

It was discovered that peaks (local maxima in latencies) were significantly better discriminators of chunks than fixed thresholds. The contrast between the approaches would be even starker, if, as would normally be the case, the number of chunks was not known *a priori* and used to set the optimal threshold. Peaks have the advantage that they use only local information about the relative magnitude of latencies to discriminate chunks. Whether the peaks method constitutes a general technique applicable beyond drawing awaits further studies in other domains.

It was found that in drawing there was a strong temporal signature of perceptual chunks in the latencies. The level of an element in the chunk hierarchy is reflected in the magnitude of the latency, the higher the level the longer the pause. The effect is sufficiently prominent to yield significant differences in individual participant data. The effect of learning is also evident in the changes in the absolute magnitude of latencies at specific chunk levels. The changes to the latencies appear to indicate when two chunks have been compiled into a single composite chunk.

These effects were consistent over the different modes of drawing, which suggests that the temporal signature reflects the structure of chunks in memory, and that the other processes of drawing, such as planning, are organized on the basis of the chunk structure. The process of drawing may magnify the effect of chunk structure rather than diminish or distort it. It seems plausible that the (sub) processes of drawing may operate in a largely serial fashion. Latencies between chunks may be longer than within chunk latencies because they encompass more sub-processes.

Further work is addressing the robustness and generalisability of the phenomena outlined in this paper. The temporal signature of chunking has been found to be apparent in other drawing domains, such as diagrammatic representations for problem solving (Lane, Cheng & Gobet, 2001).

Acknowledgements

The work was supported by the U.K. Economic and Social Research Council through the Centre for Research in Development, Instruction and Training. We are grateful to members of CREDIT for all their useful comments during the course of the studies.

References

Akin, O. (1986). *Psychology of Architectural Design*. London, Pion Ltd.

Card, S. K, Moran, T. P., & Newell, A. (1983). *The Psychology of Human-Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Chase, W., & Simon, H. (1973). Perception in Chess. *Cognitive Psychology* 4,55-81.

Egan, D. E., & Schwartz, B. J. (1979). Chunking in the recall of symbolic drawings. *Memory and Cognition*, 7(2), 149-158.

Gobet, F. (1998). Expert Chess Memory: Revisiting the Chunking Hypothesis. *Memory*, 6(3), 225-255.

Gobet, F., & Simon, H (1998). Pattern Recognition makes search possible: Comments on Holding. *Psychological Research*, 61, 204-208.

Goodnow, J., & Levine, R. (1973). The Grammar in Action: Sequence and syntax in Children's Copying. *Cognitive Psychology*, 4, 82-98.

Holden, D. H. (1992). Theories of chess skill. *Psychological Review*, 54, 10-16.

Lane, P. C. R, Cheng, P. C-H., & Gobet, F. (2001). Learning Perceptual Chunks for Problem Decomposition. In *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society* (this volume)

McLean, R., & Gregg, L. (1967). Effects of Induced chunking on Temporal Aspects of Serial Recitation. *Journal of Experimental Psychology*, 74(4), 455-459.

Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review*, 63, 81-97.

Palmer, S. (1977). Hierarchical Structure in Perceptual Representation. *Cognitive Psychology*, 9,441-474.

Reitman, J. (1976). Skilled perception in Go: Deducing Memory Structures from Inter-Response Times. *Cognitive Psychology*, 8, 357-381.

Suwa, M., & Tversky, B. (1996). What architects see in their sketches: Implications for design tools. *CHI'96-Human Factors in computing systems*, Vancouver, British Columbia, Canada, ACM.

Terrance, H. (1991). Chunking During Serial Learning by a Pigeon: I. Basic Evidence. *Journal of Experimental Psychology: Animal Behaviour Process*, 17(1), 81-93.

Tulving, E. (1962). Subjective organization in free recall of unrelated words. *Psychological Review*, 69, 344-354

Ullman, D., Wood, S., & Craig, S. (1990). The Importance of Drawing in the Mechanical Design Process. *Computer & Graphics*, 14(2), 263-274.

Van Sommers, P. (1984). *Drawing and Cognition*. New York, Cambridge University Press.

Vicente, K. (1988). Adapting the memory recall paradigm to evaluate interfaces. *Acta Psychologica*, 69, 249-278.

Wickelgren, W. (1964). Size of rehearsal group and short-term memory. *Journal of Experimental Psychology*, 68,413-419.

Wickens, C. D. (1993). *Engineering psychology and human performance*, 2nd ed. New York: HarperCollins

Modeling Tonality: Applications to Music Cognition

Elaine Chew (eniale@alum.mit.edu)
University of Southern California
Integrated Media Systems Center, and
Department of Industrial and Systems Engineering
Los Angeles, CA 90089-1450 USA

Abstract

Processing musical information is a task many of us perform effortlessly, and often, unconsciously. In order to gain a better understanding of this basic human cognitive ability, we propose a mathematical model for tonality, the underlying principles for tonal music. The model simultaneously incorporates pitch, interval, chord and key relations. It generates spatial counterparts for these musical entities by aggregating musical information. The model also serves as a framework on which to design algorithms that can mimic the human ability to organize musical input. One such skill is the ability to determine the key of a musical passage. This is equivalent to being able to pick out the most stable pitch in the passage, also known as "doh" in *solfege*. We propose a computational algorithm that mimics this human ability, and compare its performance to previous models. The algorithm is shown to predict the correct key with high accuracy. The proposed computational model serves as a research and pedagogical tool for putting forth and testing hypotheses about human perception and cognition in music. By designing efficient algorithms that mimic human cognitive abilities, we gain a better understanding of what it is that the human mind can do.

Introduction

Music cognition is a complex task requiring the integration of information at many different levels. Nevertheless, processing musical information is an act with which we are all familiar. The mind is so adept at organizing and extracting meaningful patterns when listening to music that we are often not even aware of what it is that we do when comprehending music. Some of this unconscious activity includes determining the tonal center[1], the rhythm, and the phrase structure of the piece.

I illustrate our unconscious ability to process music by a short anecdote from my own experiences. In my first semester as a pianolab[2] instructor at MIT, I encountered a few students who had no prior musical background. I asked one such student, after he carefully traced out the melodic line for Yankee Doodle, "What is the key[3] of this piece?" He responded with a reasonable question: "What do you mean by key?" I began singing the piece and stopped mid-stream. I then asked the student if he could sing me the note on which the piece should end. Without hesitation, he sang the correct pitch[4], thereby successfully picking out the first degree, and most stable pitch, in the key. The success of this method raised more questions than it answered. What is it we know that causes us to hear one pitch as being more stable than others? How does the mind assess the function of this stable pitch over time as the music evolves?

Before we can study music cognition, we first need a representation for musical structure. In this paper, we propose a mathematical model for tonality, the underlying principles of tonal music. According to Bamberger (2000), "tonality and its internal logic frame the coherence among pitch relations in the music with which [we] are most familiar." The model uses spatial proximity to represent perceived distances between musical entities. The model simultaneously incorporates representations for pitch, interval, chord and key relations.

Using this model, we design a computational algorithm to mimic human decisions in determining keys. The process of key-finding precedes the evaluation of melodic and harmonic structure, and is a fundamental problem in music cognition. We relate this new representation to previous models by Longuet-Higgins & Steedman (1971) and Krumhansl & Schmuckler (1986). The computational algorithm is shown to identify keys at a high level of accuracy, and its performance is compared to that of the two previous models.

The Representation

Western tonal music is governed by a system of rules called *tonality*. The first part of the paper proposes a geometric representation, the Spiral Array model, that captures this system of relations among tonal elements. The Spiral Array model offers a parsimonious description of the inter-relations among tonal elements, and suggests

[1] The *tonal center*, also called the *tonic* of the key, is the pitch that attains greatest stability in a musical passage.

[2] A keyboard skills class for students enrolled in Music Fundamentals and Composition courses.

[3] Excerpted from the Oxford Dictionary of Music: A *key* implies adherence, in any passage, to the note-material of one of the major or minor scales. When the pitches in a scale are ordered, the first degree of the scale gives the scale its name. This is also the most stable pitch, known as the *tonic*.

[4] A *pitch* is a sound of some frequency. High frequency sounds produce a high pitch, and low frequency sounds produce a low pitch. This is distinct from a *note*, which is a symbol that represents two properties, pitch and duration.

new ways to re-conceptualize and reorganize musical information. A hierarchical model, the Spiral Array generates representations for pitches, intervals, chords and keys within a single spatial framework, thus allowing comparisons among elements from different hierarchical levels. The basic idea behind the Spiral Array is the representation of higher level tonal elements as aggregates of their lower level components.

Spatial analogues of physical and psychological phenomena are known to be powerful tools for solving abstract intellectual problems (Shepard, 1982). Some have argued that problems in music perception can be reduced to that of finding an optimal data representation (Tanguine, 1993). Shepard (1982) determined that "the cognitive representation of musical pitch must have properties of great regularity, symmetry, and transformational invariance." The model placed all twelve chromatic pitches equally over one full turn of a spiral, and highlighted pitch height relations. Further extensions to a double helix emphasized perfect fifth interval[5] relations, but did not account for major and minor third relations.

Applying multi-dimensional scaling techniques to experimental data, Krumhansl (1978,1990) mapped listener ratings of perceived relationships between probe tones and their contexts into space. The resulting cone (1978) places pitches in the tonic triad closest to each other, confirming the psychological importance of fifth and third interval relations, which form triads. Parncutt (1988) has also presented a psychoacoustical basis for the perception of triadic units.

Another representation that incorporates spatial counterparts for both perfect fifth and major/minor third relations is the *tonnetz*, otherwise known as the Harmonic Network. This model has been used by music theorists since Riemann (see, for example, Lewin, 1987; Cohn, 1998), who posited that tonality derives from the establishing of significant tonal relationships through chord functions. Cohn (1998) has traced the earliest version of this network to the 18th century mathematician Euler, and used the *tonnetz* representation to characterize different compositional styles, focussing on preferred chord transitions in the development sections. More recently, Krumhansl (1998) presented experimental support for the psychological reality of these neo-Riemannian transformations.

Our proposed Spiral Array model derives from a three-dimensional realization of the Harmonic Network, and takes into account the inherent spiral structure of the pitch relations. It is distinct from the Harmonic Network in that it assigns spatial representations for higher level musical entities in the same structure. The representations for intervals, chords and keys are constructed as mathematical aggregates of spatial representations of their component parts.

Like the models derived from multi-dimensional scaling, the Spiral Array model uses proximity to incorporate information about perceived relationships between tonal elements. Distances between tonal entities as represented spatially in the model correspond to perceived distances among sounding entities. Perceptually close intervals are defined following the principles of music theory. In accordance with the Harmonic Network, the Spiral Array assigns greatest prominence to perfect fifth and major/minor third interval relations, placing elements related by these intervals in proximity to each other.

In the calibration of the model, the parameter values that affect proximity relations are prescribed based on a few perceived relations among pitches, intervals, chords and keys. These proximity relations will be described in a later section.

The Spiral Array Model

As the name suggests, in the Spiral Array Model, pitches are represented by points on a spiral. Adjacent pitches are related by intervals of perfect fifths. Pitches are indexed by their number of perfect fifths from C, which has been chosen arbitrarily as the reference pitch. For example, D has index two because C to G is a perfect fifth, and G to D is another. $\mathbf{P}(\mathrm{k})$ denotes the point on the spiral representing a pitch of index k. Each pitch can be defined in terms of transformations from its previous neighbor - a rotation, and a vertical translation.

$$\mathbf{P}(\mathrm{k} \quad 1) \stackrel{\text{def}}{=} \mathbf{R}\ \mathbf{P}(\mathrm{k}) \quad \mathbf{h},$$

$$\text{where } \mathbf{R} = \begin{matrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{matrix} \quad \text{and } \mathbf{h} = \begin{matrix} 0 \\ 0 \\ \mathrm{h} \end{matrix}.$$

The pitch C is arbitrarity set at the point [0,1,0].

Since the spiral makes one full turn every four pitches to line up vertically above the starting pitch position. Positions representing pitches four indices, or a major third, apart are related by a simple vertical translation:

$$\mathbf{P}(\mathrm{k} \quad 4) = \mathbf{P}(\mathrm{k}) \quad 4\ \mathbf{h}.$$

For example, C and E are a major third apart, and E is positioned vertically above C.

At this point, we diverge from the original *tonnetz* to define chord and key representations in the three-dimensional model. The added complexity of the three-dimensional realization allows one to define representations off the grid, and *within* the spiral. A chord is the composite result, or effect, of its component pitches. A key is the effect of its defining chords. We propose that this effect can be represented spatially by a convex combination of its components.

[5] Excerpted from the Oxford Dictionary of Music, an *interval* is the distance between any two pitches expressed by a number. For example, C to G is a 5th, because if we proceed up the major scale of C, the fifth pitch is G. The 4th, 5th and octave are all called Perfect. The other intervals, measured from the first pitch, in the ascending major scale are all called Major. Any Major interval can be chromatically reduced by a semitone (distance of a half step) to become Minor. If any Perfect or Minor interval is so reduced, it becomes Diminished; if any Perfect or Major interval is increased by a semitone it becomes Augmented.

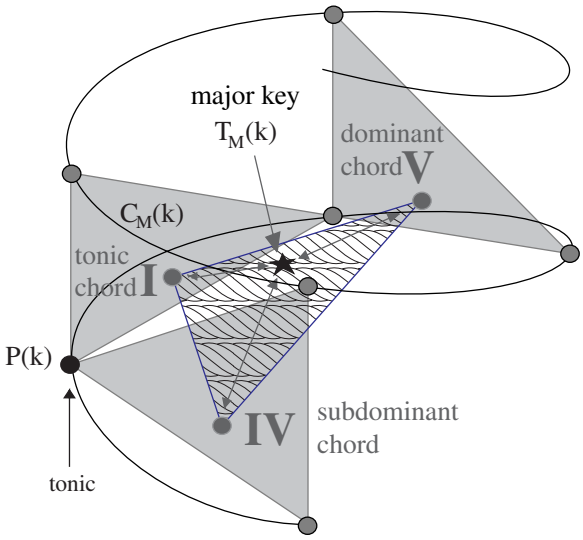


Figure 1: The Spiral Array Model.

Mathematically, the chord's representation is generated by a convex combination of its three component pitch positions. Geometrically, the chord representation resides strictly within the boundaries of the triangle outlined by the triad (see Figure 1). A chord is represented by a weighted average of its component pitch positions: the root **P**(k), the fifth **P**(k+1), and the third **P**(k+4) for major triads, and **P**(k−3) for minor triads:

The representation for a major triad is

$$\mathbf{C}_{\mathrm{M}}(\mathrm{k}) \stackrel{\text{def}}{=} w_1\ \mathbf{P}(\mathrm{k}) \quad w_2\ \mathbf{P}(\mathrm{k}\quad 1) \quad w_3\ \mathbf{P}(\mathrm{k}\quad 4),$$

$$\text{where} \qquad w_1 \quad w_2 \quad w_3 > 0 \text{ and } \sum_{i=1}^{3} w_i = 1.$$

The minor triad is generated by a similar combination,

$$\mathbf{C}_{\mathrm{m}}(\mathrm{k}) \stackrel{\text{def}}{=} u_1\ \mathbf{P}(\mathrm{k}) \quad u_2\ \mathbf{P}(\mathrm{k}\quad 1) \quad u_3\ \mathbf{P}(\mathrm{k}-3),$$

$$\text{where} \qquad u_1 \quad u_2 \quad u_3 > 0 \text{ and } \sum_{i=1}^{3} u_i = 1.$$

The weights, w_i and u_i, on the pitch positions represent the importance of the pitch to the generated chord. For longstanding psychological, physical and theoretical reasons, the root is deemed the most important, followed by the fifth, then the third. Correspondingly, the weights are constrained to be monotonically decreasing from the root, to the fifth, to the third. In order that spatial distance mirrors these relations, there are additional constraints on the aspect ratio h/r. These constraints are described in Chew (2000).

An important property of the Spiral Array is that representations of pitches in a given key occupy a compact neighborhood. Each major chord, together with its right and left neighbor major chords, combine to produce the effect of a major key. In music terminology, these chords are given names, with respect to the key, that reflect their function. The center chord is called the tonic chord (I)[6], the one to its right the dominant (V), and the one to its left the subdominant (IV). Hence, we represent the major key as a combination of its I, V and IV chords. For example, the representation of the C major key is generated by the C major, G major and F major chord representations. See Figure 1 for an example of a major key representation.

Mathematically, the representation for a major key, $\mathbf{T}_{\mathrm{M}}(\mathrm{k})$ is the weighted average of its tonic triad ($\mathbf{C}_{\mathrm{M}}(\mathrm{k})$), dominant triad ($\mathbf{C}_{\mathrm{M}}(\mathrm{k+1})$) and subdominant triad ($\mathbf{C}_{\mathrm{M}}(\mathrm{k}-1)$) representations. As before, the design objective is to have the weights correspond to each chord's significance in the key. Hence, the I chord is given the largest weight, followed by that of the V chord, then the IV chord:

$$\mathbf{T}_{\mathrm{M}}(\mathrm{k}) \stackrel{\text{def}}{=} \omega_1\ \mathbf{C}_{\mathrm{M}}(\mathrm{k}) \quad \omega_2\ \mathbf{C}_{\mathrm{M}}(\mathrm{k}\quad 1) \quad \omega_3\ \mathbf{C}_{\mathrm{M}}(\mathrm{k}-1),$$

$$\text{where} \qquad \omega_1 \quad \omega_2 \quad \omega_3 > 0 \text{ and } \sum_{i=1}^{3} \omega_i = 1.$$

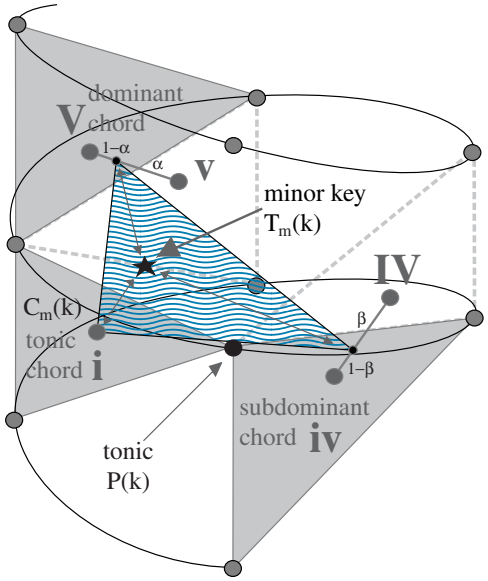


Figure 2: Geometric representation of a minor key, a composite of its tonic (i), dominants (V/v) and subdominant (iv/IV) chords.

The definition for the minor key is more complicated,

[6]We shall use roman numerals to denote chord function within a key. The number indicates the scale degree of the chord's root. For example, "I" represents the tonic chord. We adopt the convention of denoting major chords by upper case roman numerals, and minor chords by lower case ones. For example, a major chord with the tonic as root is "I" but a minor chord with the same root is "i".

but we will not go into the details at this time. It suffices to say that the center of effect for the minor key $\mathbf{T}_m(k)$ is modeled as a combination of the tonic $\mathbf{C}_m(k)$, the major and minor dominant triads $\mathbf{C}_M(k+1)$ and $\mathbf{C}_m(k+1)$, and the major and minor subdominant triad $\mathbf{C}_m(k-1)$ and $\mathbf{C}_M(k-1)$:

$$\begin{aligned}\mathbf{T}_m(k) \stackrel{\text{def}}{=} & \quad v_1 \ \mathbf{C}_m(k) \\ & v_2 \ \alpha \ \mathbf{C}_M(k \quad 1) \quad (1-\alpha) \ \mathbf{C}_m(k \quad 1) \\ & v_3 \ \beta \ \mathbf{C}_m(k-1) \quad (1-\beta) \ \mathbf{C}_M(k-1) \ , \\ \text{where} \quad & v_1 \quad v_2 \quad v_3 > 0 \text{ and } v_1 \quad v_2 \quad v_3 = 1, \\ \text{and} \quad & 0 \quad \alpha \quad 1, 0 \quad \beta \quad 1.\end{aligned}$$

See Figure 2 for the spatial representation of a minor key.

Properties of the Spiral Array Model

In the Spiral Array model, musical information is condensed, or aggregated, and represented by a single point. Proximity in the Spiral Array indicates some combination of the following: shared pitches, shared intervals, or tonal elements within a perfect fifth, major third or minor third interval of each other. This section summarizes the criteria for selecting the weights defined in the previous section so that relations between represented tonal entities have direct counterparts in the geometric structure. Details are given in Chew (2000). The criteria are summarized as follows:

1. Perceptually close intervals should be represented by shorter inter-pitch distances. For example, the closest distance between any two pitch positions denotes a perfect fifth relation; and, pitches a third apart are closer than those a second apart, etc.

2. Each chord representation is closest to its root, followed by the fifth, then the third; and, no other pitches are as close to the major chord center as its three constituent pitches.

3. The average position of two pitches an interval of a half step apart should be closest to the key related to the upper pitch; and, the average position of two pitches an interval of a perfect fourth apart should be closest to the key related to the upper pitch.

These preliminary criteria are subjective, and are by no means comprehensive. We found, through experiments, that by satisfying these few conditions, the model performed well when applied to the problem of key-finding (as described in the next sections). Thus, this could be reason to believe that with a few simple conditions, we might be able to capture the salient features in musical information in a way that concurs with listener perceptions.

Finding the Key of a Melody

The Spiral Array provided a framework on which to design viable and efficient computational algorithms for problems in the analysis and manipulation of musical information. Because the model condenses musical information to a spatial point, it allows for efficient and dynamic tracking of a streams of musical signals. Using the model, an algorithm is designed to determine the key of musical passages. We illustrate how the algorithm works by an example, "Simple Gifts". This algorithm is shown to perform better than existing ones when applied to the 24 fugue subjects in Book I of Bach's "Well-Tempered Clavier" (henceforth, referred to as the WTC). This algorithm exemplifies the concept of mapping musical information onto the Spiral Array.

Analyzing the key of a melody poses many challenges. Given a melody, one must make informed decisions about its key based on little information. Furthermore, there could be more than one equally valid answer, in which case a list for the most likely candidates for key would be more appropriate than one definite key. This section introduces the key-finding algorithm (CEG) based on the Spiral Array that returns a ranked list of possible keys. CEG is an acronym for *Center of Effect Generator*. The CEG algorithm is fundamental to the Spiral Array model and uses the model to reframe the problem of key recognition as a computationally simple one of finding a distance-minimizing representation.

In the Spiral Array, the collection of pitches in a given key defines a compact space. As pitches in a melody are sounded in sequence, the geometric shape defined by the pitch positions becomes increasingly more complex. Instead of using this complex shape to identify the key, *the algorithm collapses the pitch information down to a single point*, the center of effect (c.e.). In this manner, the pitches combine to create an object in space – a point which is the composite sum of the pitch positions.

Since keys are also defined as points in space, it is then simple to compute the distance between the c.e. and the key, and nearby keys, to determine which key is closest to the c.e. Thus the mathematical sum of pitches affords parsimonious descriptions of, and comparisons between, different pitch collections.

However, the CEG algorithm more than simply compares pitch collections. By definition, the key representations favor triadic pitch configurations, and also tonic-dominant and tonic-subdominant relationships. These representations incorporate different levels of hierarchical structure and relationships. Not all pitches are weighted equally; and, the key representation is a structured but nonlinear combination of its pitch collection. By comparing the c.e.'s to these key representations, we expect certain pitch relations to prevail.

An Example

The algorithm is best explained by an example. Consider the Shaker tune, used in Copland's symphonic suite "Appalachian Spring" (1945), shown in Figure 3.

At any point in time, the CEG method generates a c.e. from the given musical information that summarizes the tonal space generated by the pitches sounded. Define a step to be a pitch event. At each step, the pitches from the

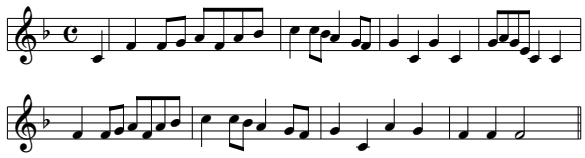


Figure 3: "Simple Gifts".

beginning to the present is weighted (multiplied) by its duration, and the c.e. is generated by *aggregating* these weighted pitch positions.

If the i-th note is represented in the Spiral Array by pitch position p_i and has duration d_i, then the *aggregate* center at the i-th pitch event is defined as:

$$c_i \stackrel{\text{def}}{=} \sum_{j=1}^{i} d_i \;\; p_i.$$

The CEG method updates its decision with each note or pitch event. The distance from the key representations to c_i is calculated and ranked. The key that is closest is ranked first, next closest second, and so on.

Figure 4 plots the exact distances from the four closest keys (F major, C major, F minor and C minor), at each successive pitch event. Observe, in the graph, that F major quickly establishes itself as the closest key. However, between pitch events $i = 22$ to 24, C major (the dominant of F) vies with F major for preeminence. The melody dwells on the dominant key area at $i = 19$ to 24, outlining the C major triad from $i = 21$ to 24. This behavior in the model concurs with listener perception.

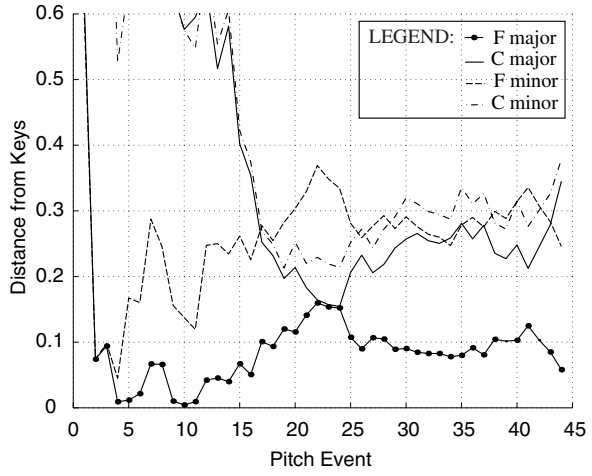


Figure 4: Distance to various keys as "Simple Gifts" unfolds.

Comparison to other Key-Finding Models

To validate the model, it was compared to Longuet-Higgins & Steedman 's Shape-Matching Algorithm (SMA) (1971) and to Krumhansl & Schmuckler's Probe Tone Profile Method (PTPM) (1986). Detailed discussions of each test run is documented in Chew (2000).

The tonic-dominant rule was devised for cases when the SMA algorithm failed to reach the desired conclusion by the end of the fugue theme. In such cases, the tonic-dominant rule derives the key from the first pitch which is assumed to be either the tonic or the dominant of the intended key. The † denotes cases when the tonic-dominant rule was applied. Numbers in brackets denote the average when considering only the fugue subjects in which the tonic-dominant rule was not applied by any of the three methods.

Table 1: Applying key-finding algorithm to Bach's fugue subjects in the WTC. (Numbers generated using h= $\overline{2\ 15}$ (r= 1), and weights across all hierarchies set to [0.516, 0.315, 0.168].)

Book I	Steps to key		
Fugue subj	CEG	PTPM	SMA
C major	2	2	16†[7]
C minor	5	5	5
C major	6	7	16
C minor	3	3	4
D major	2	2	15†
D minor	3	3	8
E major	2	6	11†
D minor	2	6	12†
E major	14	12†	11
E minor	3	2	7†
F major	4	10	6
F minor	3	15	4†
F major	3	2	8
F minor	7	18	5†
G major	2	2	15
G minor	3	3	4
A major	3	2	7†
G minor	5	5	5
A major	2	4	7
A minor	5	5	5
B major	4	4	14
B minor	2	3	6†
B major	2	11	11
B minor	3	3	7
Average	3.75 (3.57)	5.25 (4.79)	8.71 (8.21)

For the fugue subjects in Book I of the WTC, the CEG required on average 3.75 pitch events, the PTPM 5.25, and the SMA 8.71 to determine the correct key. Given a melody, a hypothesis of its key based on its first pitch is not a meaningful one. The reliability of a hypothesis based on two pitch events is still questionable. Hence, on average, the absolute minimum number of pitches re-

quired to form an opinion of the key is 3. The CEG algorithm required, on average, 3.75 steps to determine the key of the 24 fugue subjects. Based on the reasons stated, we claimed that the key-finding algorithm using the Spiral Array has an average performance that is close to optimal.

Comments

The approach detailed in this paper is computational, and mimics the manifestation of human music cognitive abilities. It proposes mathematical ways to aggregate and organize musical information. However, this does not imply that the computational algorithm describes how the human mind processes musical information. The fact that it performs well suggests that it should be considered as a method of modeling human cognition in music.

A computational model serves as a research and pedagogical tool for putting forth and testing hypotheses about human perception and cognition in music. For example, one can generate hypotheses about how humans perceive musical groupings, and implement this theory using the model.

In the melodies used for model validation, and in the "Simple Gifts" example, the aggregate points (c.e.'s) were generated cumulatively as the melodies unfolded. For lengthier examples, some decay of the information over time should be incorporated into the c.e.'s. This would be a way to model short-term memory in listening to music.

At present, the model ignores the dimension of pitch height. Clearly, pitches from different registers will generate different perceptions of relatedness. Future modifications could take into consideration the modeling of pitch height by weighting pitches from different registers differently.

The CEG algorithm currently proceeds sequentially forward through time, and cumulatively aggregates the information to produce a representation for the c.e. Cognitively, a human listener makes judgements about the key not only sequentially forward in time as the melody unfolds, but also retroactively after having gained some future information. A harmonic analysis algorithm proposed by Winograd (1968) proceeds backwards from the end of the piece; and, Temperley's (1999) extension of the Krumhansl-Schmuckler model employs dynamic programming, which also works backwards algorithmically. Future extensions of the Spiral Array Model might incorporate elements of both forward and retroactive decision-making.

By designing efficient algorithms that mimic human cognitive abilities, we gain a better understanding of what it is that the human mind can do. By studying the shortcomings of the algorithms, we can modify them, and in so doing, learn about the extent of human cognitive abilities. In the examples we discussed, the information was processed sequentially forward through time. In actual fact, the human listener can often retroactively change his or her decision about structural properties of the piece after having listened to more of the music. Attempts to model this would yield further insight as to the temporal nature of music cognition.

Acknowledgments

Jeanne Bamberger's cogent advice and unflagging support has made this research possible. This work was funded in part by the Josephine de Karman Dissertation Fellowship administered through the Massachusetts Institute of Technology.

References

Bamberger, Jeanne (2000). *Developing Musical Intuition*. New York, NY: Oxford University Press.

Chew, Elaine (2000). *Towards a Mathematical Model of Tonality*. Doctoral dissertation, Department of Operations Research, Massachusetts Institute of Technology, Cambridge, MA.

Cohn, Richard (1998). Introduction to Neo-Riemannian Theory: A Survey and a Historical Perspective. *Journal of Music Theory*, *42* (2), 167–180.

Cohn, Richard (1997). Neo-Riemannian Operations, Parsimonious Trichords, and their Tonnetz Representations. *Journal of Music Theory*, *41* (1), 1–66.

Longuet-Higgins, H. C. & Steedman, M. J. (1971). On Interpreting Bach. *Machine Intelligence*, *6*, 221.

Krumhansl, Carol L. (1990). *Cognitive Foundations of Musical Pitch*. New York, NY: Oxford University Press.

Krumhansl, C. L. & Schmuckler, M. A. (1986). The Petroushka chord: A perceptual investigation. *Music Perception*, *4*, 153–184.

Krumhansl, C. L. (1998). Perceived Triad Distance: Evidence Supporting the Psychological Reality of Neo-Riemannian Transformations. *Journal of Music Theory*, *42* (2), 265–281.

Krumhansl, C. L. (1978). *The Psychological Representation of Musical Pitch in a Tonal Context*. Doctoral dissertation, Department of Psychology, Stanford University, Stanford, CA.

Lewin, David (1987). *Generalized Musical Intervals and Transformations*. New Haven, CT: Yale University Press.

Shepard, Roger N. (1982). Structural representations of musical pitch. In D. Deutsch (Ed.), *The Psychology of Music*. New York, NY: Academic Press.

Tanguine, Andranick S. (1993). *Artificial Perception and Music Recognition*. Lecture Notes in Artificial Intelligence. New York, NY: Springer-Verlag.

Temperley, David (1999). What's Key for Key? The Krumhansl-Schmuckler Key-Finding Algorithm Reconsidered. *Music Perception*, *17* (1), 65–100

Winograd, Terry (1968). Linguistics and the Computer Analysis of Tonal Harmony. *Journal of Music Theory*, *12* (1), 2–49.

Causal Information as a Constraint on Similarity

Jessica M. Choplin (choplin@psych.ucla.edu)
Patricia W. Cheng (cheng@psych.ucla.edu)
Keith J. Holyoak (holyoak@psych.ucla.edu)
Department of Psychology
University of California, Los Angeles
405 Hilgard Ave.
Los Angeles, CA 90095-1563

Abstract

Considerable evidence indicates that causal information provides a vital constraint on conceptual representation and coherence. We investigated the role of causal information as a constraint on similarity, exploiting an asymmetry between predictive causal reasoning (given the cause, predict the effect) and diagnostic causal reasoning (given the effect, diagnose the cause). This asymmetry allowed us to isolate the effects of causal understanding from the effects of sharing non-causal features. We found that judged similarity between two objects that are identical except for one feature was affected by whether that feature was a competing potential cause of an effect or an effect of a common cause.

Causation Constrains Similarity

Any two objects have indefinitely many features in common. For example, Murphy and Medin (1985) pointed out that the number of features that plums and lawn mowers have in common is, in principle, infinite. Both weigh less than 1000 kg, and both are found on earth, in the solar system. Both cannot hear well, have an odor, are used by people, not by elephants, and so on. But despite these shared features people do not generally consider plums and lawnmowers to be similar. Intuitively, the features plums and lawnmowers have in common are not considered important. But what features are important? Why are some features important and not others? There must be criteria for constraining the sheer number of these features (Goodman, 1972; Medin, Goldstone, & Gentner, 1993). Previous researchers (e.g., Murphy & Medin, 1985) have suggested that features will be considered more important when they are diagnostic of causal function and part of a larger explanatory framework. Whether or not a feature is causal serves as a criterion by which people select important features and separate them from unimportant ones (Sloman, Love, & Ahn, 1998).

We propose a new paradigm to investigate the influence of causal knowledge on similarity and categorization. This paradigm provides an optimal way of equating the number of common and distinctive features between two objects across conditions: Equality is ensured because the stimuli were in fact identical across conditions. We use this paradigm to investigate the influence of causal knowledge as a constraint on similarity.

Experiment

To avoid confounding the effects of causal understanding with the effects of shared non-causal features, we utilized an asymmetry tested by Waldmann and Holyoak (1992). When a feature is identified as a cause, other potential causes are discounted (Morris & Larrick, 1995). Other potential causes (presented at the same time) are redundant to explain the effect and are assumed not to be causes after all. For example, if one notices a bowling ball moving, knowing that it was kicked is enough to explain the motion. There is no need to postulate another cause. By contrast, when a feature is identified as an effect, other potential effects are not discounted. For example, if one were to kick a bowling ball, the person might move the bowling ball and hurt his or her foot at the same time. The cause (the kick) would have two effects (the moving of the bowling ball and the pain in the foot). One would not assume the ball did not move simply because of the pain in the foot. The significance of this asymmetry for our current project is that it can be used to distinguish the influence of causal understanding from that of mere associations.

Two tasks were used in this experiment: first a training task and subsequently a similarity rating task. The purpose of the training task was to allow participants to learn a competing cause of an effect or an effect of a common cause. It used a two-phase blocking paradigm similar to that employed by Waldmann and Holyoak (1992). We trained participants by showing them the keys presented in Table 1. Half of the participants (predictive condition: identify features that predict an effect) were told that some of the keys could open any safe (target event in predictive condition), and they were asked to identify which features were responsible for this special ability. The other half of the participants (diagnostic condition: diagnose whether features were caused by the carving process) were told that some of the keys were carved by a special carving process (target event in diagnostic condition), and they

were asked to identify whether each feature was caused by the special carving process.

Participants saw two series of keys. In the first series, the keys were all missing a section. In the second series, this section was restored. The restoration provided an enabling condition (provided a location on which a feature could be located) for the existence of a redundant new feature that could be causally related to the target event. In the predictive condition, because old features were sufficient to explain the opening of safes, we predicted that this new, redundant feature would not be thought responsible for opening safes. By contrast, in the diagnostic condition, because a cause can have multiple effects, the new feature could very well be the result of the special carving process. We predicted that the carving process would be assumed to produce the redundant feature. Non-causal features would not produce this asymmetry.

We predicted that this causal knowledge would affect whether features are considered important in a similarity judgment. Consider a similarity judgment between a key that had the restored new feature and one that did not. If the old feature was sufficient to open safes, the key without the restored feature could open safes just as well as the key with this feature, the causal relations would not be altered, and the restored feature would not be important. By contrast, when the restored feature is the result of the carving process, the key with the restored feature received different treatment from the key without it, the causal relations would be altered, and the feature would be important.

Method

Participants

The participants were 298 students enrolled in an introductory psychology class at the University of California, Los Angeles, and 106 travelers waiting to board airplanes at Los Angeles International Airport (LAX). The students participated as part of a class exercise. The travelers participated voluntarily. Of these 404 participants, 5 failed to answer critical questions, leaving 399 responses for analysis.

Materials

As shown in Table 1, we created pictures of keys with features attached. To parallel Waldmann and Holyoak's (1992) disease symptoms, these keys had four features that we manipulated during the training phase as follows. One feature was constantly present (the t-shape on the left, called Cue C for *Constant*) and was not correlated with the target event. Another feature was present half of the time (the b-shape on the middle left, called Cue U for *Uncorrelated*) and also was not correlated with the target event. The L-shape on the middle right (called Cue P for *Predictor*) was a perfect deterministic predictor of the target event. The k-shape on the right (called Cue R for *Redundant*) was not presented at all during the first series and was only presented in the presence of Cue P during the second series. Within the second series, Cue R was a perfect predictor of the target event but only provided redundant information after Cue P.

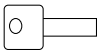
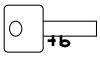
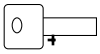
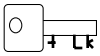
1st Series (missing section)	2nd Series (restored section)	Target Event Occurs?
		
		No
		Yes
		No
		Yes

Table 1: Key stimuli. The top key is a blank used as an initial illustration for participants. Cues C (t-shape on the left) and U (b-shape on the middle left) were not correlated with the target event. Cues P (L-shape on the middle right) and R (k-shape on the right) were correlated with the target event, but Cue R only provided redundant information.

We prepared two three-page booklets, one booklet for the predictive condition and another for the diagnostic condition. The first page was used to present the first series of keys and to elicit the causal inferences that participants first made. The second page was used to present the restored feature in the second series of keys and to elicit the causal inferences participants made after this item was restored. The third page was used to collect similarity judgments.

At the top of the first page, participants were told that some of the keys they would see either could open any safe (predictive condition) or were caused by a special carving process (diagnostic condition). It was necessary to explain why Cue R was missing from this first page. To do so, participants in both conditions were told that the keys for the first page were carved from blanks with a missing notch that looked like the first key in the left column of Table 1. In the middle of the booklets' first page a two-column table was used to present the keys. The left-hand column was labeled "Keys" and presented the four keys in the bottom four rows of the left column in Table 1. The right-hand column was used to indicate whether the target event occurred for each key.

In the predictive condition, the right-hand column asked, "Will this key open any safe?" In the diagnostic condition, it asked, "Is this key carved by the special process?" The word "Yes" was used to indicate that the target event occurred and the word "No" was used to

indicate that the target event did not occur. At the bottom of the first page, participants were asked questions on their understanding of the causal relations. In the predictive condition, participants indicated whether they thought each of the three features, Cues C, U, and P respectively, caused safes to open by circling the word "yes" or the word "no." In the diagnostic condition, they likewise indicated whether they thought the carving process caused each of the three features.

The second page resembled the first, except that the keys on the middle column of Table 1 replaced those on the left. The first key in the middle column of Table 1 replaced the first key in the left-hand column as the blank. The keys in the bottom four rows on the right replaced the keys in the bottom four rows on the left as the keys causally related to the target event. Participants were asked to indicate whether they thought each of the four features, Cues C, U, P, and R, respectively, was causally related to the target event.

The final page was used to collect similarity judgments. Seven-point scales were provided, with "1" labeled "not at all similar;" and "7" labeled "very similar." Participants compared a key containing all four features to a key that lacked Cue R. In the predictive condition, since Cue P would be sufficient to open safes, judged similarity between the two keys would be high— both keys could open safes. By contrast, in the diagnostic condition, since Cue R was caused by the carving process, judged similarity between the two keys would be low—the keys received different treatments. Participants likewise judged the similarity between a key containing all four features to one that lacked Cue P. This similarity serves as a baseline for evaluating the importance of Cue R, to reduce noise in the analyses.

Procedure

The introductory psychology students received our materials within a larger packet of materials. The packets were randomly assigned and completed within the classroom. The travelers were approached by the experimenter and asked to participate. After agreeing to participate, participants were randomly given a booklet. Participants were given pencils and instructed to follow the instructions inside the booklet. After participants had finished, the experimenter returned to collect the booklets.

Results and Discussion

The results of the training task mirrored the basic findings of Waldmann and Holyoak (1992). We divided participants into those who showed a coherent pattern of responses that indicates blocking (judged Cues C, U and R non-causal and Cue P causal) and those who showed other patterns. A blocking pattern of responses was observed for 12.5% of the participants in the predictive condition and only 4.7% of the participants in the diagnostic condition. This difference was significant, X^2 (1, n=399) = 6.13, $p < .025$. Only a minority of the participants in both conditions, however, showed this pattern. Questioning participants afterwards revealed that many based their judgments solely on the information provided on the second page, the page on which P and R both predicted the target event Many thought the second page represented a separate updated scenario. This misunderstanding likely explains why the majority of participants in both conditions judged Cue R to be causal.

To measure the importance of Cue R (relative to Cue P) in similarity judgments, we subtracted judged similarity between a key containing all four features and one that lacked Cue R from judged similarity between a key containing all four features and one that lacked Cue P. Positive scores indicated that Cue P was more important than Cue R and vice versa. As predicted, participants (n = 198) in the predictive condition judged the relative importance of Cue P to Cue R (M = 1.45) higher than participants (n = 201) in the diagnostic condition (M = 0.99), $t(397) = 2.27$, $p < .025$.

Because the size of this effect was relatively small and large numbers of participants misunderstood the task, we analyzed the relative importance of Cue P and Cue R conditionalized on causal response patterns. We divided participants into 3 groups: those who showed a coherent pattern of responses that indicates blocking (judged Cues C, U, and R non-causal and Cue P causal), those who showed a coherent pattern of responses that indicates "no blocking" (judged Cues C and U non-causal and Cues P and R causal), and those who showed other patterns. Of the participants in the predictive condition, those who showed blocking (n = 22) judged the relative importance of Cue P to Cue R (M = 2.64) greater than those who showed "no blocking" (n = 118, M = 1.28), $t(138) = 2.71$, $p < .01$. [Note: Of the participants in the diagnostic condition, the very few (n = 9) who showed blocking also judged the relative importance of Cue P to Cue R (M = 2.33) greater than those who showed "no blocking" (n = 134, M = 0.80), $t(141) = 2.81$, $p < .01$.] Most important for our hypothesis, those participants in the predictive condition who showed the blocking pattern judged the relative importance of Cue P to Cue R (M = 2.64) much higher than participants in the diagnostic condition (M = 0.99), $t(221) = 3.89$, $p < .001$, and even higher yet than participants in the diagnostic condition who did not show blocking (M = 0.8), $t(154) = 4.83$, $p < .0001$. We also replicated these results in another experiment (not reported here). It is notable that we obtained these results from a simple paper-pencil task with participants who were probably unmotivated or unable to pay close attention to the instructions or answer the questions carefully. It is likely that the differences will be enhanced under more controlled conditions, providing a means for future

studies to isolate the effects of causal understanding from the effects of mere associations.

Conclusions

The importance of causal information in category formation and conceptual coherence has been emphasized by a number of researchers (e.g., Lien & Cheng, 2000; Murphy & Medin, 1985). The results reported here extend this research to similarity judgments among novel stimuli and provide a paradigm to study this effect. A two-phase blocking paradigm such as the one used by Waldmann and Holyoak (1992) can be used to isolate the effects of causal understanding from the effects of mere association. We found that different causal schemas applied to the same correlational data affected rated similarity of novel stimuli. Of course, causality is not the only mechanism that constrains similarity. Analogical processes and higher-order relations are also responsible for providing constraints (Medin, Goldstone, & Gentner, 1993).

Why should causal information be so important? People need to discover causal relations because only such relations allow the prediction of the consequences of interventions. This constraint makes discovery of causal features vital for understanding the world and for predicting future events. Having a notion of similarity that reflects the causal relations in the world is critical for inductive inference.

Acknowledgments

We thank John Hummel, Michael Waldmann and Douglas Medin for helpful conversations and Askim Okten for running participants. This research was supported by NSF grants SBR-9729726 and SBR-9729023.

References

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and projects.* New York: Bobbs-Merrill.

Lien, Y., & Cheng, P. W. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology*, *40*, 87-137.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*, 254-278.

Morris, W. M. & Larrick, R. (1995). When one cause casts doubts on another: A normative analysis of discounting in causal attribution. *Psychological Review*, *102*, 331-355.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289-316.

Sloman, S. A., Love, B. C., & Ahn, W. (1998). Feature centrality and conceptual coherence. *Cognitive Science*, *22*, 189-228.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: asymmetries in cue competition. *Journal of Experimental Psychology: General*. *121*, 222-236.

Hemispheric lateralisation of the word length effect in Chinese character recognition

Yu-Ju Chou (sallyc@cogsci.ed.ac.uk)
Richard Shillcock (rcs@cogsci.ed.ac.uk)
Division of Informatics, University of Edinburgh,
2 Buccleuch Place, Edinburgh, EH8 9LW UK

Abstract

In the last decade, researchers of hemispheric superiority have become increasingly interested in the length effect in word recognition in alphabetic languages. But little has been known about ideographic languages like traditional Chinese. The primary aim of this study is to investigate hemispheric laterality and the word length effect in Chinese script recognition. Different-length words consisting of two-, three- and five-characters were presented unilaterally in a lexical decision task. The results, from 23 Taiwanese subjects, supported the word-length effect showing significantly different recognition latencies for the multi-character words of different length, but no significant hemispheric lateralisation. There was a significant interaction between gender and visual field, with males tending to show a right visual field advantage.

Previous studies have demonstrated hemispheric lateralisation effects in recognizing words of different length (Ellis & Young, 1985), concrete and abstract words (e.g., Ellis & Shepherd, 1974) and a word/number difference (eg, Besner, Daniels & Slade, 1982). The principle finding of a right visual field (RVF) superiority has been repeatedly reported in physically long English words: increasing word length affects the left visual field (LVF) but not the RVF presentations, resulting in a RVF superiority.

The Chinese writing system, the so-called ideogram, is distinctive from the English writing system and is supposed to present more pictorial characteristics, involving an LVF superiority in recognition tasks. Chinese stimuli presented in the LVF are hypothesized to consume shorter time in lexical decision than those presented in RVF, because the right hemisphere, directly connected to LVF, is dominant in processing pictorial images.

In 1994, Fang conducted experiments with different-length Chinese words but failed to find a significant interaction between Visual Field and Word Length. Either a significant Word Length effect or a Visual Field difference was found in separate experiments. This failure to find an interaction between length and visual field is important given the robustness of the effect in English. Below we report a replication of Fang's experiment, but with an added manipulation of gender, to investigate word recognition in Chinese.

Experiment

The primary aim of this study is to investigate hemispheric laterality and length effects in Chinese script recognition. Different-length words consisting of two-, three- and five-character were presented unilaterally in a lexical decision task.

Subjects

In this experiment, we used subjects who were able to read Chinese in traditional fonts. Twenty-three Taiwanese students studying in the University of Edinburgh volunteered. Their average age was twenty-nine for twelve males and twenty-seven for eleven females. All of them were native Mandarin speakers and were skilled Chinese readers with normal or corrected-to-normal vision. Only one of them was ambidextrous, the rest were right-handed according to self-report. The criterion of handedness was which hand they use most frequently for writing, holding chopsticks and badminton rackets, and whether there were any of their family members who were ambidextrous or left-handed.

Design

The stimuli were 120 different-length vertically displayed Chinese words and non-words consisting of 2, 3 and 5 Chinese characters, which were chosen from the Corpus of Journal Chinese (1993). Each length category contained 20 non-words and 20 real-words. This was a within-subjects repeated measures design. Half of the words were presented in the LVF and the other half the RVF. Twenty-three volunteers were divided into two groups randomly. The stimuli were arranged in a Latin Square design, therefore the words used in the first group were identical to those in second group, except that the stimuli presented in the LVF for the first group were presented in the RVF for the second group.

Stimuli

Switching the positions of two characters within one word was the way we produced the non-words. For example, three-character words like 荷包蛋 changed to 荷蛋包. The fixation point was a 4 mm × 4 mm cross (Font: Bodoni MT Ultra Bold. Size: 24. Duration 1000 msec) presented at the center of the monitor. It was to draw participants' attention and fixate their eyes on the center. A masking pattern was produced by overlapping dozens of Chinese characters that did not appear in the formal experiment. In addition, there were fifteen practice trials preceding the experiment.

All of the Chinese materials were produced by PhotoShop, and presented by Psyscope Version 1.2b5 (1994) and a Macintosh computer. The size of each character was 13 mm × 13 mm and the inter-character space was 9 mm, thus the three different lengths of words were 13 mm × 35 mm, 13 mm × 57 mm and 13 mm × 100 mm respectively. All the stimuli were presented on the screen either 2 mm to the right hand side of the fixation point or 2 mm to the left hand side of the fixation point. The smallest visual angle was equivalent to 0.25 degree from the fixation point.

Procedure

Subjects were asked to complete the personal data questionnaire before doing the experiment. After being instructed to sit in front of the computer, they were to face the center of the monitor at a distance of 450 mm to 550 mm from their eyes to the monitor, and to press the right or left button with the right or left index finger to make lexical decisions. For all the subjects, pressing the right button with the right index finger was for real-words and pressing the left button with the left index finger was for non-words. The experimenter explained the instructions and watched subjects' responses during fifteen practice trials, then subjects would be left alone while the formal experiment was progressing. The Psyscope software recorded response latencies with millisecond precision.

The sequence of presentation was firstly a fixation point, presented centrally for 1000 msec, followed by a unilaterally presented vertical Chinese stimulus which was ended by the critical response or which ended automatically after 2000 msec, followed by a masking picture presented for 1000 msec.

Analysis and results

An analysis of variance of response latencies was carried out with Visual Field and Word Length as within-subject factors and Gender as a between-subject factor. A significant main effect was found for Word Length ($F(2,42)=270.832$, $p< .001$), but was not for Gender ($F(1,21)=.429$, $p> .05$) or Visual Field ($F(1,21)=.423$, $p> .05$). In the LVF, five-character words were recognized less efficiently than both two- and three-character words, however, the differences between two- and three-character words did not reach significance. On the other hand, in the RVF, response times to 2-character words were shorter than those to 3-character words which were in turn shorter than those to 5 character words. Thus, there was a strong main effect of word length, in the predicted direction, but not of Visual Field.

The two-way interaction between Gender and Visual Field was significant ($F(1,21)=6.014$, $p< .05$), but not the one between Gender and Word Length ($F(2,42)= .898$, $p> .05$) or between Visual Field and Word Length ($F(2,42)=.250$, $p> .05$). Figure 1 shows that Females tended to recognize Chinese scripts faster than Males when scripts were presented in the LVF. But the response time difference did not reach the significance level of .05. Gender differences were not significant either in the RVF or in the LVF. On the other hand, Visual Field was marginally significant in Males ($p= .086$) but not in Females ($p= .121$). That means, for Males, words presented in the RVF tended to be better recognized than those in the LVF. There was no significant three-way interaction between Gender, Visual Field and Length ($p> .05$).

In summary, a significant main effect was found for word length, with longer words taking predictably more time to process than shorter words. Although the response times in the LVF were slightly slower than those in the RVF, the main effect of Visual Field did not reach statistical significance, either by subjects or by items, nor was there any significant interaction

Table 1. The response latency of Lexical Decision for 2-, 3- and 5-character Chinese words in different visual fields. This table presents the figures analyzed by items.

RT(msec)	RVF	LVF	For Entire Population
2-character words	780.8219	767.8909	774.3564
3-character words	863.2826	875.1924	869.2375
5-character words	1145.5332	1182.9296	1164.2314
			Mean 935.9418

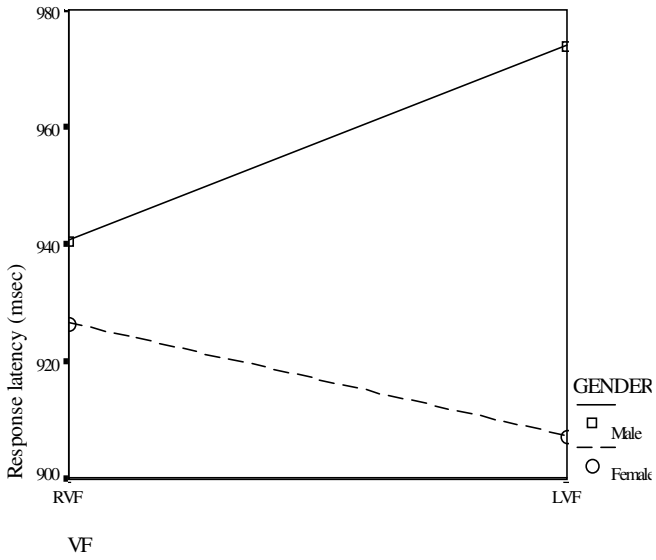


Figure 1 The interaction of Visual Field and Gender. The Visual Field difference was marginally significant in Males. That is, for Males, words presented in the RVF tended to be better recognized than those in the LVF.

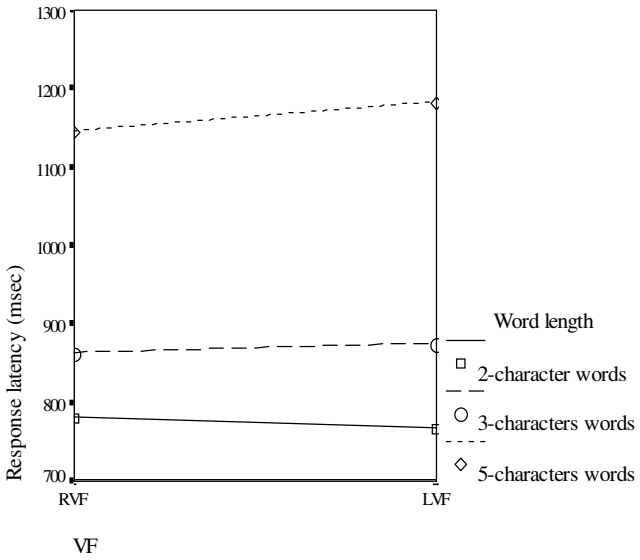


Figure 2 A significant main effect was found for Word Length, but was not for Visual Field or the interaction between Visual Field and Word Length.

between Word Length and Visual Field. In conclusion, hemispheric superiority for word length, as found in English, does not appear to exist in Chinese, from the results of this experiment.

Discussion

The methodology of this experiment was taken from Fang (1994). We extended the research of Fang by adding five-character words and introduced the gender differences. In the results, only the main effect of Word Length reached significance; neither the main effect of Visual Field nor the interaction between Visual Field and Word Length reached significance. There was, however, a significant interaction between Gender and Visual Field, the slowest reaction times coming from males responding to the LVF, supporting the idea that there are indeed visual field differences to be found in the processing of Chinese script, and comparable to those found in reading English orthography. Because of their interaction with gender, these visual field effects would seem to involve the lateralisation of phonological *versus* spatial processing. We may conclude that there is a qualitative difference between text processing in English and Chinese that prevents the emergence of a significant interaction between word length and hemifield in Chinese. Speculatively, the principal difference present in the current experiment is the vertical presentation of the words, compared with the exclusively horizontal presentation found in the relevant English experiments.

Two caveats are also necessary. Bole (1995) argued that the severity of the criteria in selecting handedness could affect the results of hemispheric superiority experiments. Mistaking left-handers for right-handers causes the data from right-handers to be less different from that of left-handers, and results in insignificant differences. Thus the current experimental procedure might be improved by using instruments such as the Edinburgh Handedness Inventory to investigate more detailed hand uses, together with a severe filter of subjects' family history in handedness.

Not all of studies of hemispheric lateralisation have reached the consistent conclusion that Chinese has significant LVF superiority with the increase of word length. But since Tzeng (1979) and following studies, it was accepted that Chinese words yield a RVF superiority whereas Chinese characters yield either a slight LVF superiority or inconsistent performance. However, many studies had failed to reduplicate these results. Thus arguments were raised from the consistency of hemispheric superiority in Chinese. Whether there is a consistent hemispheric lateralisation in recognizing Chinese words still remains doubtable.

Acknowledgments

I would like to thank my supervisor Dr. Richard Shillcock for his guidance and thank Dr. Sheng-Ping Fang for her generosity in providing the corpus and essential research papers.

References

Besner, D., Daniels, S., and Slade, C. (1982). Ideogram reading and right hemisphere language. British Journal of Psychology, 73, 21-28.

Boles, D. B. (1995). Parameters of the Bilateral effect. In Kitterle, F. L. (Eds.), Hemispheric Communication- Mechanisms and Models, UK: Lawrence Erlbaurn.

Chinese knowledge information processing group (1993). The most frequent nouns in Journal Chinese and their classification: Corpus-based research series no. 1~ 4. Taiwan: Institute of information science Academia Sinica and Institute of History & Philology Academia Sinica.

Ellis, H. D. & Shepherd, J. W., (1974). Recognition of abstract and concrete words presented in the left and right Visual Fields. Journal of Experimental Psychology, 103, 1035-1036.

Fang, S. P., (1994).English word length effects and the Chinese character-word difference: Turth or Myth? Chinese Journal of Psychology, 36(1), 59-80.

Fang, S. P., (1997). Morphological properties and the Chinese character-word difference in laterality patterns. Journal of Experimental Psychology: Human Perception and performance, 23(5), 1439-1453.

Tzeng, O. J. L., Hung, D. L., Cotton, B. and Wang, S. Y. (1979). Visual lateralization in reading Chinese characters. Nature (London), 382, 499-501.

Young A.W. & Ellis, A. W. (1985). Different Methods of Lexical access for words presented in the left and right visual hemifields. Brain and Language, 24, 326-358.

Integrating Distributional, Prosodic and Phonological Information in a Connectionist Model of Language Acquisition

Morten H. Christiansen†‡ (morten@siu.edu)
Rick A.C. Dale† (racdale@siu.edu)
†Department of Psychology
‡Department of Linguistics
Carbondale, IL 62901 USA

Abstract

Children acquire the syntactic structure of their native language with remarkable speed and reliability. Recent work in developmental psycholinguistics suggests that children may bootstrap grammatical categories and basic syntactic structure by exploiting distributional, phonological, and prosodic cues. However, these cues are probabilistic, and are individually unreliable. In this paper, we present a series of simulations exploring the integration of multiple probabilistic cues in a connectionist model. The first simulation demonstrates that multiple-cue integration promotes significantly better, faster, and more uniform acquisition of syntax. In a second simulation, we show how this model can also accommodate recent data concerning the sensitivity of young children to prosody and grammatical function words. Our third simulation illuminates the potential contribution of prenatal language experience to the acquisition of syntax through multiple-cue integration. Finally, we demonstrate the robustness of the multiple-cue model in the face of potentially distracting cues, uncorrelated with grammatical structure.

Introduction

Before children can ride a bicycle or tie their shoes, they have learned a great deal about how words are combined to form complex sentences. This achievement is especially impressive because children acquire most of this syntactic knowledge with little or no direct instruction. Nevertheless, mastering natural language syntax may be among the most difficult learning tasks that children face. In adulthood, syntactic knowledge can be characterized by constraints governing the relationship between grammatical categories of words (such as noun and verb) in a sentence. But acquiring this knowledge presents the child with a "chicken-and-egg" problem: the syntactic constraints presuppose the grammatical categories in terms of which they are defined; and the validity of grammatical categories depends on how far they support syntactic constraints. A similar "bootstrapping" problem faces a student learning an academic subject such as physics: understanding momentum or force presupposes some understanding of the physical laws in which they figure, yet these laws presuppose these very concepts. But the bootstrapping problem solved by young children seems vastly more challenging, both because the constraints governing natural language are so intricate, and because young children do not have the intellectual capacity or explicit instruction available to the academic student. Determining how children accomplish the astonishing feat of language acquisition remains a key question in cognitive science.

By 12 months, infants are attuned to the phonological and prosodic regularities of their native language (Jusczyk, 1997; Kuhl, 1999). This perceptual attunement may provide an essential scaffolding for later learning by biasing children toward aspects of the input that are particularly informative for acquiring grammatical information. Specifically, we hypothesize that integrating multiple probabilistic cues (phonological, prosodic and distributional) by perceptually attuned general-purpose learning mechanisms may hold the key to how children solve the bootstrapping problem. Multiple cues can provide reliable evidence about linguistic structure that is unavailable from any single source of information.

In the remainder of this paper, we first review empirical evidence suggesting that infants may use a combination of distributional, phonological and prosodic cues to bootstrap into language. We then report a series of simulations, demonstrating the efficacy of multiple-cue integration within a connectionist framework. Simulation 1 shows how multiple-cue integration results in better, faster and more uniform learning. Simulation 2 establishes that the trained three-cue networks are able to mimic the effect of grammatical and prosodic manipulations in a sentence comprehension study with 2-year-olds (Shady & Gerken, 1999). Simulation 3 reveals how prenatal exposure to gross-level phonological and prosodic input facilitates postnatal learning within the multiple-cue integration framework. Finally, Simulation 4 demonstrates that adding additional distracting cues, irrelevant to the syntactic acquisition task, does not hinder learning.

Cues Available for Syntax Acquisition

Although some kind of *innate* knowledge may play a role in language acquisition, it cannot solve the bootstrapping problem. Even with built-in abstract knowledge about grammatical categories and syntactic rules (e.g., Pinker, 1984), the bootstrapping problem remains formidable: children must map the right sound strings onto the right grammatical categories while determining the specific syntactic relations between these categories in their native language. Moreover, the item-specific nature of early syntactic productions challenges the usefulness of hypothesized innate grammatical categories

(Tomasello, 2000).

Language-external information may substantially contribute to language acquisition. Correlations between environmental observations relating prior semantic categories (e.g., objects and actions) and grammatical categories (e.g., nouns and verbs) may furnish a "semantic bootstrapping" solution (Pinker, 1984). However, given that children acquire linguistic distinctions with no semantic basis (e.g., gender in French, Karmiloff-Smith, 1979), semantics cannot be the only source of information involved in solving the bootstrapping problem. Another extra-linguistic factor is cultural learning where children may imitate the pairing of linguistic forms and their conventional communicative functions (Tomasello, 2000). Nonetheless, to break down the linguistic forms into relevant units, it appears that cultural learning must be coupled with language-internal learning. Moreover, because the nature of language-external and innate knowledge is difficult to assess, it is unclear how this knowledge could be quantified: There are no computational models of how such knowledge might be applied to learning basic grammatical structure.

Though perhaps not the only source of information involved in bootstrapping the child into language, the potential contribution of *language-internal* information is more readily quantified. Our test of the multiple-cue hypothesis therefore focuses on the degree to which language-internal information (phonological, prosodic and distributional) may contribute to solving the bootstrapping problem.

Phonological information—including stress, vowel quality, and duration—may help distinguish grammatical function words (e.g., determiners, prepositions, and conjunctions) from content words (nouns, verbs, adjectives, and adverbs) in English (e.g., Cutler, 1993). Phonological information may also help distinguish between nouns and verbs. For example, nouns tend to be longer than verbs in English—a difference that even 3-year-olds are sensitive to (Cassidy & Kelly, 1991). These and other phonological cues, such as differences in stress placement in multi-syllabic words, have also been found to exist cross-linguistically (see Kelly, 1992, for a review).

Prosodic information provides cues for word and phrasal/clausal segmentation and may help uncover syntactic structure (e.g., Morgan, 1996). Acoustic analyses suggest that differences in pause length, vowel duration, and pitch indicate phrase boundaries in both English and Japanese child-directed speech (Fisher & Tokura, 1996). Infants seem highly sensitive to such language-specific prosodic patterns (for reviews, see e.g., Jusczyk, 1997; Morgan, 1996)—a sensitivity that may start in utero (Mehler et al., 1988). Prosodic information also improves sentence comprehension in two-year-olds (Shady & Gerken, 1999). Results from an artificial language learning experiment with adults show that prosodic marking of syntactic phrase boundaries facilitates learning (Morgan, Meier & Newport, 1987). Unfortunately, prosody is partly affected by a number of non-syntactic factors, such as breathing patterns (Fernald & McRoberts, 1996), resulting in an imperfect mapping between prosody and syntax. Nonetheless, infants' sensitivity to prosody provides a rich potential source of syntactic information (Morgan, 1996).

None of these cues in isolation suffice to solve the bootstrapping problem; rather, they must be integrated to overcome the partial reliability of individual cues. Previous connectionist simulations by Christiansen, Allen and Seidenberg (1998) have pointed to efficient and robust learning methods for multiple-cue integration in speech segmentation. Integration of phonological (lexical stress), prosodic (utterance boundary), and distributional (phonetic segment sequences) information resulted in reliable segmentation, outperforming the use of individual cues. The efficacy of multiple-cue integration has also been confirmed in artificial language learning experiments (e.g., McDonald & Plauche, 1995).

By one year, children's perceptual attunement is likely to allow them to utilize language-internal probabilistic cues (for reviews, see e.g., Jusczyk, 1997; Kuhl, 1999). For example, infants appear sensitive to the acoustic differences between function and content words (Shi, Werker & Morgan, 1999) and the relationship between function words and prosody in speech (Shafer, Shucard, Shucard & Gerken, 1998). Young infants can detect differences in syllable number among isolated words (Bijeljac, Bertoncini & Mehler, 1993)—a possible cue to noun/verb differences. Moreover, infants are accomplished distributional learners (e.g., Saffran, Aslin & Newport, 1996), and importantly, they are capable of multiple-cue integration (Mattys, Jusczyk, Luce & Morgan, 1999). When solving the bootstrapping problem children are also likely to benefit from specific properties of child-directed speech, such as the predominance of short sentences (Newport, Gleitman & Gleitman, 1977) and the cross-linguistically more robust prosody (Kuhl et al., 1997).

This review has indicated the range of language-internal cues available for language acquisition, that these cues affect learning and processing, and that mechanisms exist for multiple-cue integration. What is yet unknown is how far these cues can be combined to solve the bootstrapping problem (Fernald & McRoberts, 1996).

Simulation 1: Multiple-Cue Integration

Although the multiple-cue approach is gaining support in developmental psycholinguistics, its computational efficacy still remains to be established. The simulations reported in this paper are therefore intended as a first step toward a computational approach to multiple-cue integration, seeking to test the potential advantages of this approach to syntactic acquisition. Based on our previous experience with modeling multiple-cue integration in speech segmentation (Christiansen et al., 1998), we used a simple recurrent network (SRN; Elman, 1990) to model the integration of multiple cues. The networks were trained on corpora of artificial child-directed speech generated by a well-motivated grammar that includes three probabilistic cues to grammatical structure: word length,

lexical stress and pitch. Simulation 1 demonstrates how the integration of these three cues benefits the acquisition of syntactic structure by comparing performance across the eight possible cue combinations.

Method

Networks Ten SRNs were used in each cue condition, with an initial weight randomization in the interval [-0.1; 0.1]. Learning rate was set to 0.1, and momentum to 0. Each input to the networks contained a localist representation of a word, and a constellation of cue units depending on its assigned cue condition. Networks were required to predict the next word in a sentence along with the corresponding cues for that word. With a total of 44 words and a pause marking boundaries between utterances, the networks had 45 input units. Networks in the condition with all available cues had an additional five input units. The number of input and output units thus varied between 45-50 across conditions. Each network had 80 hidden units and 80 context units.

Materials We constructed a complex grammar based on independent analyses of child-directed corpora (Bernstein-Ratner, 1984; Korman, 1984), and a study of child-directed speech by mother-daughter pairs (Fisher & Tokura, 1996). As illustrated in Table 1, the grammar included three primary sentence types: declarative, imperative, and interrogative sentences. Each type consisted of a variety of common utterances reflecting the child's exposure. For example, declarative sentences most frequently appeared as transitive or intransitive verb constructions (*the boy chases the cat*, *the boy swims*), but also included predication using *be* (*the horse is pretty*) and second person pronominal constructions commonly found in child-directed corpora (*you are a boy*). Interrogative sentences were composed of wh-questions (*where are the boys?*, *where do the boys swim?*), and questions formed by using auxiliary verbs (*do the boys walk?*, *are the cats pretty?*). Imperatives were the simplest class of sentences, appearing as intransitive or transitive verb phrases (*kiss the bunny*, *sleep*). Subject-verb agreement was upheld in the grammar, along with appropriate determiners accompanying nouns (*the cars* vs. **a cars*).

Two basic cues were available to all networks. The fundamental distributional information inherent in the grammar could be exploited by all networks in this experiment. As a second basic cue, utterance-boundary pauses signalled grammatically distinct utterances with 92% reliability (Broen, 1972). This was encoded as a single unit that was activated at the end of all but 8% of the sentences. Other semi-reliable prosodic and phonological cues accompanied the phrase-structure grammar: word length, stress, and pitch. Network groups were constructed using different combinations of these three cues. Cassidy and Kelly (1991) demonstrated that syllable count is a cue available to English speakers to distinguish nouns and verbs. They found that the probability of a single syllable word to be a noun rather than a verb is 38%. This probability rises to 76% at two syllables, and 92% at three. We selected verb and noun tokens that exhibited this distinction, whereas the length of the remaining words were typical for their class (i.e., function words tended to be monosyllabic). Word length was represented in terms of three units using thermometer encoding—that is, one unit would be on for monosyllabic words, two for bisyllabic words, and three for trisyllabic words. Pitch change is a cue associated with syllables that precede pauses. Fisher and Tokura (1996) found that these pauses signalled grammatically distinct utterances with 96% accuracy in child-directed speech, allowing pitch to serve as a cue to grammatical structure. In the networks, this cue was a single unit that would be activated at the final word in an utterance. Finally, we used a single unit to encode lexical stress as a possible cue to distinguish stressed content words from the reduced, unstressed form of function words. This unit would be on for all content words.

Table 1: **The Stochastic Phrase Structure Grammar Used to Generate Training Corpora**

S → Imperative [0.1] | Interrogative [0.3] | Declarative [0.6]
Declarative → NP VP [0.7] | NP-ADJ [0.1] | That-NP [0.075] | You-P [0.125]
NP-ADJ → NP is/are adjective
That-NP → that/those is/are NP
You-P → you are NP
Imperative → VP
Interrogative → Wh-Question [0.65] | Aux-Question [0.35]
Wh-Question → where/who/what is/are NP [0.5] | where/who/what do/does NP VP [0.5]
Aux-Question → do/does NP VP [0.33] | do/does NP wanna VP [0.33] | is/are NP adjective [0.34]
NP → a/the N-sing/N-plur
VP → V-int | V-trans NP

Procedure Eight groups of networks, one for each combination of cues, were trained on corpora consisting of 10,000 sentences generated from the grammar. Each network within a group was trained on a different training corpus. Training consisted of 200,000 input/output presentations (words), or approximately 5 passes through the training corpus. Each group of networks had cues added to its training corpus depending on cue condition. Networks were expected to predict the next word in a sentence, along with the appropriate cue values. A corpus consisting of 1,000 novel sentences was generated for testing. Performance was measured by assessing the networks' ability to predict the next set of grammatical items given prior context—and, importantly, this measure did not include predictions of cue information.

To provide a statistical benchmark with which to compare network performance, we "trained" bigram and trigram models on the same corpora as the networks. These finite-state models, borrowed from computational linguistics, provide a simple prediction method based on strings of two (bigrams) or three (trigrams) consecutive

words. Comparisons with these simple models provide an indication of whether the networks are learning more than simple two- or three-word associations.

Results

All networks achieved better performance than the standard bigram/trigram models ($p's < .0001$), suggesting that the networks had acquired knowledge of syntactic structure beyond the information associated with simple pairs or triples of words. The nets provided with phonological/prosodic cues achieved significantly better performance than base networks ($p's < .02$). Using trigram performance as criterion, all multiple-cue networks surpassed this level of performance faster than the base networks ($p's < .002$). Moreover, the three-cue networks were significantly faster than the single-cue networks ($p's < .001$). Finally, using Brown-Forsyth tests for variability in the final level of performance, we found that the three-cue networks also exhibited significantly more uniform learning than the base networks ($F(1,18) = 5.14, p < .04$).

Simulation 2: Sentence Comprehension in Two-Year-Olds

Simulation 1 provides evidence for the general feasibility of the multiple-cue integration approach. However, to further strengthen the model's credibility closer contact with relevant human data is needed. In the current simulation, we demonstrate that the three-cue networks from Simulation 1 are able to accommodate recent data showing that two-year-olds can integrate grammatical markers (function words) and prosodic cues in sentence comprehension (Shady & Gerken, 1999: Expt. 1). In this study, children heard sentences, such as (1), in one of three prosodic conditions depending on pause location: early natural [e], late natural [l], and unnatural [u]. Each sentence moreover involved one of three grammatical markers: grammatical (the), ungrammatical (was), and nonsense (gub).

1. Find [e] the/was/gub [u] dog [l] for me.

The child's task was to identify the correct picture corresponding to the target noun (dog). Simulation 2 replicates this by using comparable stimuli, and assessing the noun activations.

Method

Networks Twelve networks from Simulation 1 were used in each prosodic condition. This number was chosen to match the number of infants in the Shady and Gerken (1999) experiment. An additional unit was added to the networks to encode the nonsense word (gub) in Shady and Gerken's experiment.

Materials We constructed a sample set of sentences from our grammar that could be modified to match the stimuli in Shady and Gerken. Twelve sentences for each prosody condition (pause location) were constructed. Pauses were represented by activating the utterance-boundary unit. Because these pauses probabilistically signal grammatically distinct utterances, the utterance-boundary unit provides a good approximation of what the children in the experiment would experience. Finally, the nonsense word was added to the stimuli for the within group condition (grammatical vs. ungrammatical vs. nonsense). Adjusting for vocabulary differences, the networks were tested on comparable sentences, such as (2):

2. Where does [e] the/is/gub [u] dog [l] eat?

Procedure Each group of networks was exposed to the set of sentences corresponding with its assigned pause location (early vs. late vs. unnatural). No learning took place, since the fully-trained networks were used. To approximate the picture selection task in the experiment, we measured the degree to which the networks would activate the groups of nouns following the/is/gub. The two conditions were expected to affect the activation of the nouns.

Results

Shady and Gerken (1999) reported a significant effect of prosody on the picture selection task. The same was true for our networks ($F(2,33) = 1,253.07, p < .0001$). The late natural condition elicited the highest noun activation, followed by the early natural condition, and with the unnatural condition yielding the least activation. The experiment also revealed an effect of grammaticality as did our networks ($F(2,70) = 69.85, p < .0001$), showing the most activation following the determiner, then for the nonsense word, and lastly for the ungrammatical word. This replication of the human data confers further support for Simulation 1 as a model of language acquisition by multiple-cue integration.

Simulation 3: The Role of Prenatal Exposure

Studies of 4-day-old infants suggest that the attunement to prosodic information may begin prior to birth (Mehler et al., 1988). We suggest that this prenatal exposure to language may provide a scaffolding for later syntactic acquisition by initially focusing learning on certain aspects of prosody and gross-level properties of phonology (such as word length) that later will play an important role in postnatal multiple-cue integration. In the current simulation, we test this hypothesis using the connectionist model from Simulations 1 and 2. If this scaffolding hypothesis is correct, we would expect that prenatal exposure corresponding to what infants receive in the womb would result in improved acquisition of syntactic structure.

Method

Networks Ten SRNs were used in both prenatal and non-prenatal groups, with the same initial conditions and training details as Simulation 1. Each network was supplied with the full range of cues used in Simulation 1.

Materials A set of "filtered" prenatal stimuli was generated using the same grammar as previously (Table 1), with the exception that input/output patterns now ignored individual words and only involved the units encoding word length, stress, pitch change and utterance boundaries. The postnatal stimuli were the same as in Simulation 1.

Procedure The networks in the prenatal group were first trained on 100,000 input/output filtered presentations drawn from a corpus of 10,000 new sentences. Following this prenatal exposure, the nets were then trained on the full input patterns exactly as in Simulation 1. The non-prenatal group only received training on the postnatal corpora. As previously, networks were required to predict the following word and corresponding cues. Performance was again measured by the prediction of following words, ignoring the cue units.

Results

Both network groups exhibited significantly higher performance than the bigram/trigram models ($F(1,18) = 25.32, p < .0001$ for prenatal, $F(1,18) = 12.03, p < .01$ for non-prenatal), again indicating that the networks are acquiring complex grammatical regularities that go beyond simple adjacency relations. We compared the performance of the two network groups across different degrees of training using a two-factor analysis of variance (ANOVA) with training condition (prenatal vs. non-prenatal) as the between-network factor and amount of training as within-network factor (five levels of training measured in 20,000 input/output presentation intervals). There was a main effect of training condition ($F(1,18) = 12.36, p < .01$), suggesting that prenatal exposure significantly improved learning. A main effect of degrees of training ($F(9,162) = 15.96, p < .001$) reveals that both network groups benefitted significantly from training. An interaction between training conditions and degrees of training indicates that the prenatal networks learned significantly better than postnatal networks ($F(1,18) = 9.90, p < 0.01$). The exposure to prenatal input—void of any information about individual words—promotes better performance on the prediction task; thus providing computational support for the prenatal scaffolding hypothesis.

Simulation 4: Multiple-Cue Integration with Useful and Distracting Cues

A possible objection to the previous simulations is that our networks succeed at multiple-cue integration because they are "hand-fed" cues that are at least partially relevant for syntax acquisition. Consequently, performance may potentially drop significantly if the networks themselves had to discover which cues were partially relevant and which are not. Simulation 4 therefore tests the robustness of our multiple-cue approach when faced with additional, uncorrelated distractor cues. Accordingly, we added three distractor cues to the previous three reliable cues. These new cues encoded the presence of word-initial vowels, word-final voicing, and relative (male/female) speaker pitch—all acoustically salient in speech, but which do not appear to cue syntactic structure.

Method

Networks Networks, groups and training details were the same as in Simulation 3, except for the addition of the three additional input units encoding the distractor cues.

Materials The three distractor cues were added to the stimuli used in Simulation 3. Two of the cues were phonetic and therefore available only in postnatal training. The word-initial vowel cue appears in all words across classes. The second distractor cue, word-final voicing, also does not provide useful distinguishing properties of word classes. Finally, as an additional prenatal and postnatal cue, overall pitch quality was added to the stimuli. This was intended to capture whether the speaker was female or male. In prenatal training, this probability was set to be extremely high (90%), and lower in postnatal training (60%). In the womb, the mother's voice naturally provides most of the input during the final trimester when the infant's auditory system has begun to function (Rubel, 1985). The probability used here intended to capture that some experience would likely derive from other speakers as well. In postnatal training this probability drops, representing exposure to male members of the linguistic community, but still favoring mother-child interactions.

Procedure Prenatal stimuli included the three previous semi-reliable cues, and only the additional prosodic, distractor cue encoding relative speaker pitch. In the postnatal stimuli, all three distractor cues were added. Training and testing details were the same as in Simulation 3.

Results

As in Simulations 1 and 3, both groups performed significantly better than the bigram/trigram models ($F(1,18) = 18.95, p < .0001$ for prenatal, and $F(1,18) = 14.27, p < .001$ for non-prenatal). We repeated the two-factor ANOVA computed for Simulation 2, revealing a main effect for training condition ($F(1,18) = 4.76, p < 0.05$) and degrees of training ($F(9,162) = 13.88, p < .0001$). This indicates that the presence of the distractor cues did not hinder the improved performance following prenatal language exposure. As in Simulation 3, the prenatal networks learned comparatively faster than the non-prenatal networks ($F(1,18) = 5.31, p < .05$).

To determine how the distractor cues may have affected performance, we compared the prenatal condition in Simulation 3 with that of the current simulation. There was no significant difference in performance across the two simulations ($F(1,18) = 0.13, p = 0.72$). A further comparison between these non-prenatal networks and the bare networks in Simulation 1 showed that the networks trained with cues of mixed reliability significantly outperformed networks trained without any cues ($F(1,18) = 14.27, p < .001$). This indicates that the

uncorrelated cues did not prevent the networks from integrating the partially reliable ones towards learning grammatical structure.

Conclusion

A growing bulk of evidence from developmental cognitive science has suggested that bootstrapping into language acquisition may be a process of integrating multiple sources of probabilistic information, each of which is individually unreliable, but jointly advantageous. However, what has so far been missing is a comprehensive demonstration of the computational feasibility of this approach. With the series of simulations reported here we have taken the first step toward establishing the computational advantages of multiple-cue integration. Simulation 1 demonstrated that providing SRNs with prosodic and phonological cues significantly improves their acquisition of syntactic structure—despite the fact that these cues are only partially reliable. The multiple-cue integration approach gains further support from Simulation 2, showing that the three-cue networks can mimic children's sensitivity to both prosodic and grammatical cues in sentence comprehension. The model also illustrates the potential value of prenatal exposure, since Simulation 3 revealed significant benefits for networks receiving such input. Finally, Simulation 4 provides evidence for the robustness of our neural network model, since highly unreliable cues did not interfere with the integration process. This implementation of our model still exhibited significant performance advantages over networks not receiving cues at all. Moreover, all the network models consistently performed better than the statistical benchmarks, the bigram and trigram models. This has important theoretical implications because it suggests that the SRNs acquired complex knowledge of grammatical structure and not merely simple two- or three-word co-occurrence statistics. Overall, the simulation results presented in this paper provide support not only for the multiple-cue integration approach in general, but also for a connectionist approach to the integration of distributional, prosodic and phonological information in language acquisition.

References

Bernstein-Ratner, N. (1984). Patterns of vowel modification in motherese. *Journal of Child Language*, *11*, 557–578.

Bijeljac, R., Bertoncini, J. & Mehler, J. (1993). How do 4-day-old infants categorize multisyllabic utterances? *Developmental Psychology*, *29*, 711–721.

Broen, P. (1972). *The verbal environment of the language-learning child*. ASHA Monographs, No. 17. Washington, DC: American Speech and Hearing Society.

Cassidy, K.W. & Kelly, M.H. (1991). Phonological information for grammatical category assignments. *Journal of Memory and Language*, *30*, 348–369.

Christiansen, M.H., Allen, J. & Seidenberg, M.S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*, 221–268.

Cutler, A. (1993). Phonological cues to open- and closed-class words in the processing of spoken sentences. *Journal of Psycholinguistic Research*, *22*, 109–131.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179-211.

Fernald, A. & McRoberts, G. (1996). Prosodic bootstrapping: A critical analysis of the argument and the evidence. In J.L Morgan & K. Demuth (Eds.),*From Signal to syntax* (pp. 365–388). Mahwah, NJ: Lawrence Erlbaum Associates.

Fisher, C. & Tokura, H. (1996). Acoustic cues to grammatical structure in infant-directed speech: Cross-linguistic evidence. *Child Development*, *67*, 3192–3218.

Jusczyk, P.W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

Karmiloff-Smith, A. (1979). *A functional approach to child language: A study of determiners and reference*. Cambridge, UK: Cambridge University Press.

Kelly, M.H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review*, *99*, 349–364.

Korman, M. (1984). Adaptive aspects of maternal vocalization in differing contexts at ten weeks. *First Language*, *5*, 44–45.

Kuhl, P.K. (1999). Speech, language, and the brain: Innate preparation for learning. In M. Konishi & M. Hauser (Eds.), *Neural mechanisms of communication* (pp. 419–450). Cambridge, MA: MIT Press.

Kuhl, P.K., Andruski, J.E., Chistovich, I.A., Chistovich, L.A., Kozhevnikova, E.V., Ryskina, V.L., Stolyarova, E.I., Sundberg, U. & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, *277*, 684–686.

Mattys, S.L., Jusczyk, P.W., Luce, P.A. & Morgan, J.L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, *38*, 465–494.

McDonald, J.L. & Plauche, M. (1995). Single and correlated cues in an artificial language learning paradigm. *Language and Speech*, *38*, 223–236.

Mehler, J., Jusczyk, P.W., Lambertz, G., Halsted, N., Bertoncini, J. & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*, 143–178.

Morgan, J.L. (1996). Prosody and the roots of parsing. *Language and Cognitive Processes*, *11*, 69–106.

Morgan, J.L., Meier, R.P. & Newport, E.L. (1987). Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Psychology*, *19*, 498–550.

Newport, E.L., Gleitman, H. & Gleitman, L.R. (1977). Mother, Id rather do it myself: Some effects and non-effects of maternal speech style. In C.E. Snow & C.A. Ferguson (Eds.), *Talking to children: Language input and acquisition* (pp. 109–149). Cambridge, UK: Cambridge University Press.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Rubel, E.W. (1985). Auditory system development. In G. Gottlieb & N.A. Krasnegor (Eds.), *Measurement of audition and vision in the first year of postnatal life*. Norwood, NJ: Ablex.

Saffran, J.R, Aslin, R.N. & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Shady, M., & Gerken, L.A. (1999). Grammatical and caregiver cues in early sentence comprehension. *Journal of Child Language*, *26*, 163–175.

Shafer, V.L., Shucard, D.W., Shucard, J.L. & Gerken, L.A. (1998). An electrophysiological study of infants' sensitivity to the sound patterns of English speech. *Journal of Speech, Language, and Hearing Research*, *41*, 874–886.

Shi, R., Werker, J.F., & Morgan, J.L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, *72*, B11–B21.

Tomasello, M. (2000). The item-based nature of children's early syntactic development. *Trends in Cognitive Sciences*, *4*, 156–163.

Using Distributional Measures to Model Typicality in Categorization

Louise Connell (louise.connell@ucd.ie)
Department of Computer Science, University College Dublin,
Belfield, Dublin 4, Ireland

Michael Ramscar (michael@cogsci.ed.ac.uk)
School of Cognitive Science, University of Edinburgh,
2 Buccleuch Place, Edinburgh, EH8 9LW, Scotland.

Abstract

Typicality effects are ordinarily tied to concepts and conceptualization. The underlying assumption in much of categorization research is that effects such as typicality are reflective of stored conceptual structure. This paper questions this assumption by simulating typicality effects by the use of a co-occurrence model of language, Latent Semantic Analysis (LSA). Despite being a statistical tool based on simple word co-occurrence, LSA successfully simulates subject data relating to typicality effects and the effects of context on categories. Moreover, it does so without any explicit coding of categories or semantic features. The model is then used to successfully predict participants' judgements of typicality in context. In the light of the findings reported here, we question the traditional interpretation of typicality data: are these data reflective of underlying structure in people's concepts, or are they reflective of the distributional properties of the linguistic environments in which they find themselves.

Introduction

The world contains a myriad of objects and events that an intelligent observer could seemingly infinitely partition and generalise from. So how it is that humans can adopt a particular partitioning in the mass of data that confronts them? How do they pick out regularities in the stuff of experience and index them using words? What are these regularities? And how do humans recognise, communicate, learn and reason with them? These questions are central to cognitive science, and traditionally, their close linkage has tempted researchers to seek a unified answer to them: categorization – the act of grouping things in the world – has been commonly linked to the representation of concepts[1], with many researchers assuming that a theory of one provides for the other (Armstrong, Gleitman & Gleitman, 1983; Keil, 1987; Lakoff, 1987).

In much of this work, it is assumed that linguistic behavior (such as naming features associated with a concept, c.f. Rosch, 1973) is determined by, and reflective of, underlying concepts that are grounded in perceptual experience of objects and artifacts themselves. Here, we wish to consider the idea that language itself is part of the environment that determines conceptual behavior. A growing body of research indicates that *distributional information* may play a powerful role in many aspects of human cognition. In particular, it has been proposed that people can exploit statistical regularities in language to accomplish a range of conceptual and perceptual learning tasks. Saffran, Aslin & Newport (1996; see also Saffran, Newport, & Aslin; 1996) have demonstrated that infants and adults are sensitive to simple conditional probability statistics, suggesting one way in which the ability to segment the speech stream into words may be realized. Redington, Chater & Finch (1998) suggest that distributional information may contribute to the acquisition of syntactic knowledge by children. MacDonald & Ramscar (this volume) have shown how information derived from a 100 million word corpus can be used to manipulate subjects' contextual experience with marginally familiar and nonce words, demonstrating that similarity judgements involving these words are affected by the distributional properties of the contexts in which they were read.

The objective of this paper is to examine the extent to which co-occurrence techniques can model human categorization data: What is the relationship between typicality judgements and distributional information? Indeed, are the responses people provide in typicality experiments more reflective of the distributional properties of the linguistic environments in which they find themselves than they are of the underlying structure of people's concepts?

Typicality Effects

The first empirical evidence of typicality effects was provided by Rosch (1973), who found participants judged some category members as more (proto)typical than others. Rosch (1973) gave subjects a category name such as *fruit* with a list of members such as apple,

[1] In the experiments reported, we follow the common assumption (Medin & Smith, 1984; Komatsu, 1992) that categories are classes, concepts are their mental representations and that an instance is a specific example of a category member.

fig, olive, plum, pineapple, strawberry, etc. and asked subjects to rate on a 7-point scale how good an example each member was of its category. The results showed a clear trend of category gradedness – apples are consistently judged a typical *fruit*, while olives are atypical. Further evidence underlines the pervasiveness of typicality (or "goodness of example") and its ability to predict a variety of results. Typicality was found to predict reaction times in sentence verification tasks (Rosch, 1973; McCloskey & Glucksberg, 1979) and order of item output when subjects are asked to name members of a category (Barsalou & Sewell, 1985).

Roth & Shoben (1983) showed that the context a concept appears in affects the typicality of its instances. A typical *bird* in the context-free sense may be a *robin*, but if it appears in the context "The bird walked across the barnyard", then *chicken* would instead be typical. Subject reaction times to sentence verification tasks are faster for the contextually appropriate item (*chicken*) than the normally typical, but contextually inappropriate item (*robin*). Roth and Shoben found that measures of typicality determined in isolation no longer play a predictive role once context has been introduced.

Typicality, Substitutability and LSA

According to Rosch (1978): "The meaning of words is intimately tied to their use in sentences. Prototypicality ratings for members of superordinate categories predict the extent to which the member term is substitutable for the superordinate word in sentences."

This notion of contextual substitutability has a parallel in distributional approaches to word meanings (e.g. Landauer & Dumais, 1997; Burgess & Lund, 1997). In a distributional model of word meaning such as Latent Semantic Analysis (LSA), the corpus analysis calculates a contextual distribution for each lexeme encountered by counting the frequency with which it co-occurs[2] with every other word in the corpus. The contextual distribution of a word can then be summarized by a vector – or point in high-dimensional space – that shows the frequency with which it is associated with the other lexemes in the corpora. In this way, two words that tend to occur in similar linguistic contexts – i.e. are *distributionally* similar – will be positioned closer together in semantic space than two words which do not share as much distributional information. By using the proximity of points in semantic space as a measure of their contextual substitutability, LSA offers a tidy metric of distributional similarity.

Rosch (1973; 1978) held that such substitutability arises as a result of similarities between the underlying structures of the concepts representing the words (although describing these underlying structures has proven elusive, see Komatsu, 1992; Ramscar & Hahn, in submission). However, distributional theories suggest that information about substitutability and word similarity can instead be gleaned from the structure of the *linguistic environment*. Such information is readily – and objectively - obtainable for the purposes of model building and hypothesis testing.

[2] How words are used together within a particular context, such as a paragraph or moving-window.

Experiment 1 – Canonical Typicality

The purpose of this experiment is to examine whether data from typicality studies (Rosch, 1973; Armstrong, Gleitman & Gleitman, 1983; Malt & Smith, 1984) can be modeled using a distributional measure. Specifically, it was predicted that subject typicality scores from previous studies would correlate with a distributional measure (LSA; Landauer & Dumais, 1997) when comparing similarity scores for category members against their superordinate category name.

Materials

Each set of typicality data was divided up according to the taxonomy used in the original study: Set A was taken from Rosch (1973), B from Armstrong, Gleitman & Gleitman (1983), and C from Malt & Smith (1984).

Within these three data sets, 18 sets of typicality ratings existed, across 12 separate categories due to overlap between categories used in Sets A, B and C.

Procedure

For each category in each data set, all items were compared in LSA to the superordinate category name and the similarity scores noted. All scores were calculated in LSA using a corpus whose texts are thought to be representative of reading materials experienced by students up to first year in college[3].

The LSA scores were then scaled from the given [-1, +1] range to fit the standard 7-point typicality scale used in the subject studies, where a score of 1 represents the most typical rating. Malt & Smith used the 7-point scale in reverse order (where 7 represented most typical) so these scores were inverted. LSA score scaling was done by aligning the highest of the LSA scores for each category with the most typical rank on the 7-point scale[4]; i.e. the highest LSA score for a category would be matched to 1, and the other scores falling proportionately towards 7.

[3] Using General Reading up to 1st Year College semantic space, with term-to-term comparison and maximum factors.

[4] The exact formula used is as follows: Where X is the LSA score one wishes to scale and M is the maximum LSA score for this category set:
Scaled LSA score = $M - (M - 1) / (M * X)$.

Results

Spearman's rank correlation (*rho*) was used to compare scaled LSA and subject scores. The global rank correlation between the subject ratings and LSA scores across Sets A, B and C (193 items) was *rho*=0.515 (2-tailed $p<0.001$). Many of the categories that failed to produce greatly significant correlations correlated significantly with the removal of one member, due to it having an extremely high or low LSA score (usually because of its low frequency of occurrence in the corpus). See Table 1 for full LSA results. Also, Set A / Set B correlations for their 4 shared categories of *sport*, *fruit*, *vehicle* and *vegetable* were *rho*=1.0 ($p<0.01$), *rho*=0.943 ($p<0.05$), *rho*=0.886 ($p<0.05$) and *rho*=0.886 ($p<0.05$) respectively.

Table 1: Rank correlation coefficients *rho* (with levels of significance *p*) between LSA and subject scores

Set	Category	Initial category	Adjusted category
A	sport	1.000 ($p<0.01$)	
	fruit	0.886 ($p<0.05$)	
	vehicle	0.829 ($p<0.10$)	1.000 ($p<0.05$)
	crime	0.814 ($p<0.10$)	0.975 ($p<$ (0.10)
	bird	0.714 ($p<0.10$)	0.900 ($p<0.10$)
	science	0.414 (-)	0.675 ($p<0.10$)
	vegetable	0.371 (-)	
B	sport	0.811 ($p<0.01$)	
	vehicle	0.788 ($p<0.01$)	
	vegetable	0.580 ($p<0.10$)	0.745 ($p<0.05$)
	fruit	0.539 ($p<0.10$)	0.748 ($p<0.05$)
	female	0.346 (-)	0.558 ($p<0.10$)
C	trees	0.705 ($p<0.01$)	
	clothing	0.521 ($p<0.05$)	0.676 ($p<0.05$)
	furniture	0.466 ($p<0.05$)	0.609 ($p<0.01$)
	bird	0.375 (-)	0.640 ($p<0.05$)
	fruit	0.157 (-)	
	flowers	-0.499 (-)	

Values (-) represent insignificant correlation of p>0.10

It must be noted that the same rank correlation coefficient results in differing levels of significance within Table 1. This is due to different sizes in categories' data sets (from 5 to 20), where the same score could be significant for one size set and not another; e.g. perfect rank correlation of 1.000 is significant to $p<0.01$ with $N=10$, but only to $p<0.05$ when $N=5$. This high sensitivity to the degrees of freedom from small-sized data sets is why one outlying item was capable of skewing the rank correlation. With small data sets such as these, the power of the tests being used is restricted and they are overly sensitive to individual data points. Larger category data sets are to be found in Sets B and C, where although the rank correlation coefficients may be lower, they are more significant. Thus, it seems reasonable to consider as marginally significant those results where $p<0.10$, given the constraints of the data.

Discussion

In this experiment, LSA similarity scores correlated significantly with subject typicality ratings. Without any hand-coding of category membership or salient features, LSA's semantic space successfully modeled gradients of typicality within categories. Significant global correlation existed between LSA-to-subject typicality ratings at *rho*=0.515 ($p<0.001$). Items that subjects judged typical correlated with those that LSA scored highly in similarity with the category name. The same correlation is true of items that subjects judged to be highly atypical members of their category – these received low similarity scores in LSA. The more closely the ranking of LSA scores mirrored that of the subjects', the higher the correlation, and the closer the level of significance dropped to zero.

Regarding the categories themselves, there were cases where LSA modeled a category's typicality gradient successfully in one data set but not in another. An example of this is the category *fruit*, which was modeled with rank correlation of *rho*=0.886 ($p<0.05$) in Set A and 0.748 ($p<0.05$) in Set B (adjusted), but failed to correlate significantly at all in Set C.

Only one of the 5 category types in Set B came from what Armstrong, Gleitman & Gleitman (1983) term as "well-defined" categories – the category *female*. Despite Armstrong, Gleitman and Gleitman's designation of this category as well-defined, it seems reasonable to regard typicality in *female* as one would any other category examined in this experiment – a measure of contextual substitutability. In this case, the contextual substitutability shown by LSA similarity scores failed to convincingly model the typicality scores for *female*, only reaching correlation of 0.558 ($p<0.10$) when the category was adjusted. We propose the reason for this is that typicality ratings for a category such as *female* are subject to social conditioning in a way other categories such as *fruit* or *sport* are not. For example, the item that LSA scored highest against *female* was *housewife*, which was next followed by *chairwoman*. Although this simply reflects the general contextual substitutability of the words across all of LSA's corpora, it also reflects a ranking that may not be found within a social group. It would be inconsistent for a group of subjects to rate *housewife* as the most typical *female* (a stereotyped sexist attitude), while rating *chairwoman* (a stereotyped politically correct attitude) closely behind. Thus LSA may have failed to convincingly model this category's typicality gradient because it reflects an average of social attitudes across

its corpora, and not just those of one particular group – 1980's Philadelphia undergraduates.

One of the most interesting findings is that in 3 out of 4 cases of shared categories between Set A and Set B, LSA provided as good a fit to Set A typicality ratings as Set B did. When the item *skis* was removed from Set A's *vehicle* category, LSA's correlation bettered that of Set B (with the sole exception of the category *vegetable*). This serves to make an important point and put the data in Table 1 into perspective: it suggests that the difference between subject groups in Rosch's (1973) and Armstrong, Gleitman & Gleitman's (1983) experiments is comparable to the difference between LSA and human subjects. In other words, a co-occurrence model like LSA is as successful at matching the typicality gradients of a subject group as another subject group would be.

Experiment 2 – Contextual Typicality

The first experiment indicates that a co-occurrence model such as LSA can be used to model typicality judgements in canonical (context-free) categories. However, categorization is also subject to linguistic context, whose capacity to skew typicality has been demonstrated by Roth & Shoben (1983).

Having examined canonical typicality in Experiment 1, the purpose of Experiment 2 was to test if LSA could be used to predict subject responses for typicality in context. The hypothesis was that LSA could predict human judgements of exemplar appropriateness (typicality) for given context sentences. LSA similarity scores for each context sentence[5] and respective category members were used to form significantly different clusters of appropriate (high scores / high similarity) and inappropriate (low scores / low similarity) items. It was predicted that subject ratings of typicality in context for these items would fall into the same clusters, and that these clusters would also be significantly different.

Materials

Materials consisted of 7 context sets, each of which consisted of a context sentence and 10 possible members of the category. 3 of the context sentences were taken from Roth & Shoben (1983), the other 4 created for this experiment. Category members were chosen in two ways, to form the appropriate and inappropriate clusters for the context.

First, appropriate items were found by randomly selecting 4-5 high-level category members (e.g. *cow* not *calf* for category *animal*) that appeared in the LSA list of 1500 near neighbors of the context sentence[6]. This list corresponds to the 1500 points in LSA's high-dimensional space that are closest to the context sentence, and would receive the highest similarity scores.

Second, inappropriate items were found by compiling a large list of category members and selecting the 5-6 of those that had the lowest (preferably negative) LSA similarity score against the context sentence.

These materials were split into two sections. Each section consisted of 7 context sets, now with 5 items, selected so that there were at least 2 of both appropriate and inappropriate items in the set and so that each category member appeared only once per section. Subjects received one section apiece, with presentation of section 1 or 2 alternated between subjects. All 35 items within each section were presented in random order, resampled for each subject.

Participants

19 native speakers of English took part in this experiment. All were volunteers who participated by completing an electronic questionnaire.

Procedure

LSA Procedure The scores were calculated in LSA by comparing the context sentence to each item in the list, using the same corpus as for Experiment 1[7].

The LSA scores were then scaled from the given [-1, +1] range to fit the standard 7-point typicality scale used in the subject studies. Due to the presence of very low negative LSA scores, this was done by aligning the extremes of the LSA scores for each category with the opposite extremes of the 7-point scale; i.e. the highest LSA score for a category would be matched to 1, the lowest score to 7, and the intermediate scores falling proportionately in between[8].

Human Procedure Participants read instructions that explained typicality and the 7-point scale as per Rosch (1973) and Armstrong, Gleitman & Gleitman (1983). They were then given this example of a context sentence (not used in experiment) "The girl played the GUITAR while the others sang around the campfire",

[5] The LSA score for a sentence is computed by taking a weighted average of the vectors for each word.

[6] The sentence was processed as a pseudodoc using maximum factors in the same semantic space as used in Experiment 1, from which all words in the corpus with a frequency of less than or equal to 5 had been removed.

[7] Using document-to-term comparison and maximum factors.

[8] The exact formula used is as follows: Where X is the LSA score one wishes to scale, M is the maximum LSA score for this category set and L is the midpoint of the LSA score range for this category set:
Scaled LSA score = $4 - [(L - X) * 3] / (L * M)$.
(4 = midpoint of 7-pt scale; 3 = scale end [7] – midpoint [4]).

and told to consider the appropriateness of the capitalized word in the context given.

Participants were asked not to spend more than 10 seconds deciding on what score to give, and were told that it would not be possible to go back and change an answer (the questionnaire was set up to prevent participants from doing this).

Results

Subjects agreed with LSA's predictions of typicality for 62 of the total 70 items – 10/10 items in 3 context sets, 9/10 items in 3 further context sets, and 5/10 in the remaining context set. Significant difference in clusters, not rank correlation, is the important factor here, because even subject data with low correlation to the LSA score may fall into the two specified clusters (and thus provide support for the main prediction).

Table 2: Wilcoxon's *W* and significance of difference between clusters for each context sentence.

Context Sentence	LSA	Subjects
Stacy volunteered to milk the *animal* whenever she visited the farm *	10 ($p<0.01$)	10 ($p<0.01$)
Fran pleaded with her father to let her ride the *animal* *	15 ($p<0.01$)	15 ($p<0.01$)
The *bird* swooped down on the helpless mouse and carried it off	10 ($p<0.01$)	10 ($p<0.01$)
Jane liked to listen to the *bird* singing in the garden	15 ($p<0.01$)	18 ($p<0.1$) *10 ($p<0.05$) adjusted*
Jimmy loved everything sweet and liked to eat a *fruit* with his lunch every day	15 ($p<0.01$)	18 ($p<0.1$) *10 ($p<0.05$) adjusted*
Sophie was a natural athlete and she enjoyed spending every day at *sport* training	15 ($p<0.01$)	19.5 ($p<0.1$) *10.5($p<0.05$) adjusted*
During the mid morning break the two secretaries gossiped as they drank the *beverage* *	15 ($p<0.01$)	25 ($p<0.7$)**

* Sentences taken from Roth & Shoben (1983)
** Not significant but included for completeness

For all 7 context sets, Mann-Whitney (Wilcoxon Summed Ranks, 2-tailed) tests showed the LSA scores fell into two significantly different clusters. When testing subject scores for difference between the predicted clusters, results varied from three context sets showing significant differences at $p<0.01$ (those at 10/10 agreement), to one set failing to achieve any significant difference at $p=0.69$ (5/10 agreement). Data for clustering in both LSA and subject scores are given in Table 2. Three of the context sets that only produced clusters which were significantly different to $p<0.10$ were those where subjects agreed with LSA-predicted clusters for 9/10 items. With the removal of this lone contentious item, each of these three adjusted subject sets achieved significance of $p<0.025$ (see Table 2).

Discussion

The results support the basic hypothesis that, in the majority of cases, distributional information (in this case modeled in LSA) can predict whether members of a category will be appropriate or inappropriate in a given context. In other words, it can predict human judgements of typicality in context as well as in canonical categories (as demonstrated in Experiment 1). For example, LSA predicted in the context set for *animal* ("Fran pleaded with her father...") that the item *elephant* would be placed in the inappropriate cluster, even though it is entirely possible to ride on an elephant.[9]

In 3 of the 7 context sets, subject typicality scores agreed with LSA predicted clusters for 10/10 items and separated the clusters to a difference significance of $p<0.01$. These sets involved natural kinds as the category for which typicality was taken (*animal*, *bird*). In a further 3 context sets, subjects agreed with LSA's clustering for 9/10 items and separated the clusters to a significant difference of $p<0.05$ when these 9 items were considered. For these sets, two categories were of natural kinds (*bird*, *fruit*) and one was an abstract artifact kind (*sport*). Finally, the context set for which only 5/10 items were agreed to be in the predicted clusters was also for an artifact kind (*beverage*). This suggests that distributional information (or at least, LSA) may perform better in predicting the contextual typicality of natural kinds than artifact kinds. This is perhaps as a result of the vectors for artifact kinds containing a greater degree of contextual variation and thus scoring more unpredictably against the context sentence. Such a theory is compatible with psychological data showing that artifact kinds are processed differently because they may be found in a variety of functional and relational roles, and/or are often polysemous (e.g. Wisniewski & Gentner, 1991).

General Discussion

The success of a distributional measure (LSA) in these modeling experiments suggests interesting possibilities for a theory of categorization based in context, that incorporates information from the structure of language as well as from the structure of the world. Distributional models of language use a representation that is learned from the language alone, assuming that the way words co-occur with one another gives rise to

[9] Although we anticipated a problem with participants' judgements here, the prediction was consistent with the data, where *elephant* received a typicality score of 4.1 and resided in the inappropriate cluster. Is this respect, LSA predictions were sometimes unexpectedly appropriate.

clues about their semantic meaning. Gleitman (1990) has discussed a similar approach with regards to first language acquisition, where this type of representation can be easily learned from an individual's response to their linguistic environment, thus lending a psychologically plausible base to such a theory.

In this respect, the results of these simulations raise interesting questions with regard to people's mental representations of the meanings of words: Do people use distributional information to construct their representation of word meanings? Or are distributional properties of words (which models such as LSA extract) merely an epiphenomenonon; a reflection of the fact that underlying concepts share certain semantic features? By the latter account, the distributional properties associated with words would arise *because* the concepts underlying the words possess certain features, and it is sensitivity to similarities between these concepts that subjects actually manifest. However, MacDonald & Ramscar (in submission) have shown that manipulating the distributional properties of the contexts in which nonce words are read can significantly influence similarity judgements between existing words and nonces. This indicates that not *all* distributional responses can be explained in terms of existing conceptual structure – nonce words won't have an existing conceptual structure. Equally, it seems highly unlikely that the structure of the linguistic environment is entirely unreflective of the structure that people extract from their interactions with the world.

What the results presented here (and other distributional research) seem to indicate is that any proper characterization of conceptual thought will have to consider more than just the information that comes from physical experience and the physical environment. One must also consider experience of language, and the structure of the linguistic environments in which speakers find themselves.

Acknowledgments

We thank Dermot Lynott and Dan Yarlett for many insightful comments on the work reported in this paper.

References

Armstrong, S. L., Gleitman, L. R. & Gleitman, H., (1983). What some concepts might not be. *Cognition, 13*, 263-308.

Barsalou, L. W. and D. R. Sewell (1985). Contrasting the representations of scripts and categories. *Journal of Memory and Language, 24*, 646-665.

Burgess, C. & Lund, K., (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes, 12,* 1-34.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition, 1,* 3-55.

Keil, F.C. (1987). Conceptual Development and Category Structure. In U. Neisser (Ed*.), Concepts and Conceptual Development: Ecological and intellectual Factors in Categorization.* Cambridge:Cambridge University Press.

Komatsu, L., (1992). Recent views of conceptual structure. *Psychological Bulletin, 112,* 500-526.

Lakoff, G., (1987). *Women, Fire and Dangerous Things.* University of Chicago Press.

Landauer, T. K. & Dumais, S. T., (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104,* 211-240.

Malt, B. & Smith, E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior, 23,* 250-269.

McCloskey, M., & Glucksberg, S. (1979). Decision processes in verifying category membership statements: Implications for models of semantic memory. *Cognitive Psychology, 11,* 1-37.

MacDonald, S & Ramscar, M. J. A. (this volume) Testing the distributional hypothesis: the influence of context on judgements of semantic similarity. *This volume*.

Medin, D. & Smith, E. (1984). Concepts and concept formation. *Annual Review of Psychology*, 35, 113-138.

Redington, M., Chater, N. & Finch, S. (1998). Distributional information: a powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*, 425-469.

Ramscar, M. J. A. & Hahn, U. (in submission). *What family resemblances are not: Categorisation and the concept of 'concept'*. Manuscript in submission

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.) *Cognitive Development and the Acquisition of Language*. New York: Academic Press.

Rosch, E., (1978). Principles of Categorization. In E. Rosch and B. B. Lloyd (Eds.), *Cognition and categorization*. Hillsdale, N.J.: Erlbaum.

Roth, E. M. & Shoben, E. J., (1983). The effect of context on the structure of categories. *Cognitive Psychology, 15,* 346-378.

Saffran, J. R., Aslin, R. N. & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926-1928.

Saffran, J. R., Newport, E. L. & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606-621

Wisniewski, E. J., & Gentner, D. (1991). On the combinatorial semantics of noun pairs: Minor and major adjustments to meaning. In G. B. Simpson (Ed.), *Understanding word and sentence*. Amsterdam: Elsevier.

Young Children's Construction of Operational Definitions in Magnetism: the role of cognitive readiness and scaffolding the learning environment

Constantinos P. Constantinou (c.p.constantinou@ucy.ac.cy), Athanassios Raftopoulos (raftop@ucy.ac.cy), George Spanoudis (spanoud@ucy.ac.cy)
Department of Educational Sciences, University of Cyprus
P.O. Box 20537, 1678 Nicosia, Cyprus.

Abstract

In this paper, we examine the importance of scaffolding the environment and the role of cognitive readiness in young childrens' construction of operational definitions in magnetism. We discuss various resource constraints and the conceptual background of preschoolers. Then we present an experimental study of 165 children aged 4-6 who took part in an extended structured intervention in which they were guided to construct two operational definitions of a magnet. The two definitions differed with regard to the cognitive demands imposed upon the children attempting to construct them. The construction of the second operational definition required cognitive abilities that the construction of the first did not. Our results demonstrate that children older than 5 years are mostly able to construct both definitions while younger children are able to construct only the first one. Based on this result, we discuss the issue of cognitive readiness and its role in learning. Additionally, by teaching one experimental group of older children the second definition directly and observing their limited success to construct it, we argue for the necessary role of scaffolding the conceptual structure of the curriculum materials to achieve learning.

Introduction

Real understanding of a concept is only demonstrated when children can construct operational definitions (McDermott, 1996). The reason may be that guiding children to formulate such definitions fosters the formation of explicit declarative knowledge, which benefits understanding of the concepts (Peters, et. al., 1999).

A child can be expected to understand those concepts for which the epistemologically prerequisite concepts are manageable and the necessary cognitive resources that will allow concept construction have been acquired. The set of these concepts cannot be determined *a priori* but only through empirical research.

A successful model of teaching should be designed with an eye to the limitations that constrain children's perception and interpretation of the world. It should also seek to take advantage of these limitations by scaffolding the learning environment in ways that enable children to explore their difficulties and to explicitly resolve them.

One of the concepts that have proven recalcitrant to successful teaching is that of magnets. Research (Barrow, 1987; Gagliari, 1981; Selman et. al., 1982) shows that preschoolers notice magnetic attraction but cannot spontaneously offer a successful definition of a magnet. It also shows that they find magnetic repulsion more difficult. Since repulsion is important in differentiating between magnets and non-magnetized ferrous materials, it is important that an understanding of magnets be based on the interactions between two magnets (Gagliari, 1981).

In this paper, we aim to examine the extent to which preschoolers (4-6 years of age) can successfully construct two operational definitions of a magnet. In the first part of the paper, we present the theoretical background regarding preschoolers' representation of the world and the constraints that shape it. We will discuss their intuitive theories of magnetism and will elaborate on the term "cognitive readiness", by which we mean a set of cognitive skills and resources at a given age.Then, we will present a school-based didactic intervention aiming to test whether preschoolers could successfully be taught two different operational definitions of a magnet. The first definition treated magnetism as a substantial property of some objects. The second definition requires that children understand magnetism as a relation between two objects. This, in turn requires, first, that children can combine information from two independent sources, and second, that they can coordinate causal schemes. In this sense, the second definition is cognitively more demanding than the first one. With this approach, we aim to explore the way in which children's cognitive readiness, or lack thereof, manifests itself in the construction of operational definitions of a magnet. Prior to and after the intervention we carried out individual interviews designed to evaluate the children's prior experience with magnets and their ability to apply each of the two definitions.

Theoretical Background

The cognitive basis of intuitive notions of magnetism

One of the basic traits of the preconceptual experiential background, is the tendency of children to believe that properties belong to objects and exist independently of interactions with other bodies. Thus they find it very difficult to comprehend physical

concepts that are relational and to represent processes among interacting bodies, such as, electrical or magnetic interactions, gravity or thermal transfer (Carey, 1986; Chi, 1993).

This trait may find its explanation in the fact that preschoolers encode and remember only categorical information (Demetriou *et. al.*, 1993; Fisher, 1980), and furthermore, that they fail to encode comparative or contrastive information. Thus, they systematically misrepresent data that contain comparative information (Thelen and Smith, 1994).

The concepts that are based upon our basic interaction with the world have a meaning constituent that is not conceptual but experiential in nature and upon which semantic content is subsequently progressively built. The experiential non-propositional meaning constituent is an image schema (Johnson, 1987). This intuitive, or primitive, meaning is so fundamental that it constitutes the "hard core" of some of our concepts. diSessa (1993) calls such meaning carriers "phenomenological primitives" and they are the main tools that render experience meaningful in the first instance. By virtue of the fact that they are grounded in experience, image schemata are very persistent. This explains the tremendous difficulties we encounter in our effort to revise an image shema, should it prove to be inconsistent with the corresponding scientific concepts. Examples of such schemata are the various image schemata of "force", and pertaining to the topic of this paper, the image of attraction.

In so far as magnetism is concerned, the image schema, or phenomenological primitive, of an attractive force gives rise to the "pulling model" (Erickson, 1994). According to this model the magnet is viewed as an object that has the capability to pull other objects, or, sticks to other objects (Barrow, 1987; Gagliari, 1981; Selman et. al., 1982). This conception of magnetism characterizes the intuitive conceptions of children up to the age of 10.

Children use a wide variety of causal explanations to account for phenomena, and seem to observe both domain-general and domain-specific causal principles. Among the domain general causal principles is the thesis that the causes must resemble their effect - "homeopathy" (Spinger and Keil 1991). With regard to magnetism, this means that it is difficult for young children to grasp a causal mechanism that can lead to disparate and often antithetical effects, as in the case of a magnet that can attract and repel objects. Demetriou et. al. (1993a) argue that children between the ages of 3 and 5 employ proto-causal schemes that allow them to differentiate causal from random sequences on the basis of the structure of events in space and time. Around the age of 5, children can coordinate the proto-causal schemes and search for causes by testing some hypothesis that can be formulated on the basis of the surface structure of the event.

Our discussion thus far reveals the cognitive basis of some of the difficulties that children have with respect to magnets. To recapitulate: (a) children view magnetism as a substantial property of some objects; (b) this property is conceived as the force magnets have in order to pull toward them, or "stick to", other objects; and (c) children find it difficult to understand the fact that a magnet can both attract and repel other objects.

Learning is conditioned upon the epistemological structure of the domain and the cognitive profile of the learner, including the resources that the learning system bears upon the task. Thus, it is important to identify these resources and examine their impact on learning.

Resource limitations and their role in learning

It is now well documented (Kemler 1983; Shepp 1978; Smith and Kemler, 1977) that in classification and discrimination tasks, younger children (up to around 4.5 years of age) tend to perform in a way which suggests that they perceive dimensional combinations as integral, and consequently base their decisions with respect to the classificatory or discriminatory tasks on the perception of the overall similarity of the presented stimuli. Older children classify or discriminate among objects by attending to, and analyzing, the dimensions of the stimuli. This allows them to perceive the embedded structure in the stimuli array. This ability is the decisive factor determining age differences with respect to performance. Further evidence (Gentner and Toupin, 1986; Vosniadou, 1987) suggests that the same maturational trend is attested in tasks of analogy, metaphor, and knowledge transfer.

Another resource limitation is the lack of capacity of children younger than 5 years to combine and integrate information from two independent sources (Halford and McDonald, 1977). They do not have what Piaget calls the capacity to perform logical multiplication. This capacity can influence children's construction of concepts whose definitions require the integration of information.

Resource limitations, far from being a hindrance to learning, render learning possible by scaffolding the environment and the information it feeds to the learning system (Elman, 1991; Raftopoulos, 1997). The key to success lies in effectively limiting the initial access of the cognizer to the full body of information, and in the gradual introduction of the system to the domain's full complexity. This "undersampling" of a complex domain gives the system the opportunity to learn first the domain's features and regularities, and eventually build on them the more complex features which will allow it to generalize.

This conforms with Clark and Thorton's (1997) account of learning, in which problems can be divided into two categories: those whose solution requires finding of the surface structure of the data, that is, first order regularities (type-1), and those whose solution requires finding the deep structure of the data, that is, the more abstract regularities, (type-2 problems). Problems

of type-1 can relatively easily be solved by means of an inductive search of the relevant problem space that can extract the basic statistical distributions in the data. Statistical procedure cannot be applied directly to type-2 problems. Thus, problems of type-2 could be solved if transformed to type-1 problems. This can be achieved by recoding and reorganizing the data so that they can render clear the underlying hidden structure. The first operational definition is a typical case of a type-1 problem. It requires that the children limit themselves to examining only information regarding the phenomenon of attraction between bodies. When the first definition is understood, the children "know" that those bodies that can attract others are to be categorized as magnets, all other factors becoming irrelevant to the problem. The property of "attraction" becomes the recoding schema on the basis of which they will attack the second definition, which is a type-2 problem. Once other factors have been eliminated, those children that have the appropriate cognitive readiness include information regarding mutual repulsion and eventually also understand the second operational definition. This is a clear case of undersampling the domain and scaffolding the environment.

In this part of the paper we have discussed certain characteristic developmental trends of preschool aged children, namely, the emergence of the ability to perform logical multiplication, the emergence of the ability to combine proto-causal schemes, and the emergence of the ability to discover embedded structure in an array and go beyond surface similarities. There is also evidence that all these skills appear around the age of 5. Thus, around the age of 5 preschoolers acquire skills that enhance their comprehension of the surrounding world. We will say that these children acquire a "cognitive readiness".

In the next section, we will present an experimental study that was designed in the light of the preceding theoretical framework and aimed to examine the way preschoolers can be quided to comprehend magnetism, by constructing operational definitions of a magnet.

The Experimental Study

The research questions of our study are the following:

(a) can preschoolers learn successfully to construct operational definitions of magnetism?
(b) can preschoolers construct a relational operational definition based on mutual attractions and repulsions, overcoming persistent epistemological obstacles?
(c) is effective scaffolding of the learning environment a necessary condition for preschoolers to construct the second, more complex, operational definition?

Children Participants

The sample included 165 children ranging in age from 3 years and 11 months to 5 years and 7 months (sample mean 4 years and 10 months and standard deviation 6 months). The children attended three kindergartens in a small city and were distributed in six different classrooms. All teachers underwent training in content knowledge, and curriculum implementation procedures.

Description of the teaching intervention

In our intervention, we explicitly encourage children to use evidence (particularly their own observations) to always support their viewpoints. The curriculum materials are very detailed in offering guidance to the teachers as to how to create an environment where children are encouraged to express themselves and every opinion is valued. Some aspects of the curriculum, such as guiding children to classify objects according to material, are not trivial and the activity sequence required many trials before it could be refined to a version that was deemed effective. The unit includes 6 sequential lessons as follows:

1. Exploring magnets
2. Metals and non-metals
3. Are all metals attracted by a magnet?
4. How can I tell if something is a magnet?
5. Magnets with other magnets
6. Is there another way to tell if something is a magnet?

Our interest was in investigating children's ability to construct and apply consistently operational definitions uniquely distinguishing a magnet from other objects. The curriculum guided children to formulate the following operational definitions:

I. Find two objects that do not attract each other. Does your object attract both of them? If yes, then it is a magnet. If not, then try with other objects. (Lesson 4)

II. Find two objects that when approached in some orientation they attract each other AND when approached in another orientation they repel each other. Both of these objects are magnets. (Lesson 6)

The words attraction, orientation and repulsion were usually avoided by the children. Instead they would typically use the words pull, another way, and push, respectively.

Data Collection

The data was collected through individual interviews prior to the intervention, at the end of lesson 4 (Operational Definition I) and in the two weeks following lesson 6 (Operational Definition II).

Task 1: Pretest Interviews

In our initial (pre-test) evaluation, each child was given a bowl with ten objects including 3 magnets and was asked to group them on the basis of interactions between objects. Children were encouraged to settle on one best classification and this was recorded, both photographically and in note form, at the end of the interview. Each classification was then coded based on the criterion that the child seemed to employ.

Task 2: Operational Definition I

In this task, children were presented with a group of 10 objects each of which was hidden in a matchbox wrapped in white paper and sealed with cellotape. This group of hidden objects included only one magnet. Children were explicitly told this and were then asked to give directions to the interviewer so that s/he could identify the magnet. The interviewer acted out the directions so that the child could see the result. All interviews were audiotaped. Children's responses were then coded as a success or a failure based on whether they could provide directions so that the interviewer could apply Operational Definition I consistently.

A set of directions was graded as successful only if it specified all of the following three items:

a) Finding two objects at random that attracted each other,
b) Testing each of the objects repeatedly with a third object;
c) Rejecting one of the two objects that was found not to interact with a third object interacting with the other of the initial two.

Operational Definition II

In this second task, children were presented with a set of 10 identically looking objects wrapped in the same manner as in the first task. They were explicitly told that the objects included two magnets this time and they were asked to give directions to find both magnets in one go. The interviewer again acted out the directions so that the child could see the result. Some children spontaneously resorted to applying operational definition I. When this happened the interviewer clarified once that they were to give one set of directions so that both magnets could be found simultaneously. A response was graded as successful only if it specified all of the following three items:

a) Finding two objects at random that attracted each other,
b) Testing different orientations of the two objects to see of they also repelled
c) Rejecting one object at a time until two objects were found that both attracted and repelled.

Any response that did not include anyone of these items was deemed unsuccessful.

Results

Pre-test interviews

Table 1 presents the results from children's responses in the initial interviews. Many noticed the magnets but ignored them in their groupings. 47 children did not recognize the magnets in their bowls. Most of the groupings were on the basis of colour, shape, heaviness or more than one of these criteria were used simultaneously. The responses of 32 children could not be categorized unambiguously and the criterion is listed as "unidentified".

Table 1: Criteria used by children (N=165) to classify objects in their initial pre-test interviews

Criterion	Number of children
Magnetic attraction	11
Shape	26
Color	32
Heaviness	19
Texture	9
Material	7
Mixed	29
Unidentified	32

118 children appeared to recognize the magnets in their bowls. This number gives an indication of how many of these children remembered having seen a magnet prior to the start of our intervention. Only 11 children noticed that there was a magnet among their objects *and* used it in any way to influence their grouping. This number provides an upper bound on the number of children who may have been able to give an acceptable form of Operational Definition I prior to the intervention. Both, these 11 children and the 32 children who used unidentified criteria were distributed roughly evenly in the 6 classrooms.

Operational Definition Tasks

The total number of children participants is N=165. The number of children who received the whole treatment (lessons 1-6) is N=136. 90.4% (N=123) of these children performed successfully on the Operational Definition I task. Only 47.8% (N=65) of these performed successfully on the Operational Definition II task. Another class of children (N=29) were only examined for Operational Definition II. The success rate for this class was 41.4%.

In order to test the hypothesis on cognitive readiness, we decided to separate the children into three different age groups. Based on our hypothesis we would expect children older than five to perform significantly better on Operational Definition II than children aged below 5. Table 2 presents the children's performance in the two operational definition tasks as a function of age (N=136). The children are divided into three groups according to age (below 4 and 6 months, above 5 and in between). The percentage of children who perform successfully on Operational Definition I is very high. This seems to suggest that after appropriate intervention virtually all children in this age range are able to construct Operational Definition I. Operational Definition II has a substantially lower success rate for every age group. Operational Definition II also demonstrates a strong dependence on age. Only 14.3% below age 4 and 6 months perform successfully. In contrast, 85.4% of children above age 5 are able to consistently construct Operational Definition II.

Table 2: Children's performance on Operational Definitions I and II for different age groups

Group Age Range	N (N_t=136)	Mean Age (yr:mos)	Stand. Devn (mos)	Success rate Opernal Defn I	Success rate Opernal Defn II
> 5	41	5:4	2	95.1%	85.4%
4.5 - 5	46	4:8	2	97.8%	50.0%
< 4.5	49	4:3	2	79.6%	14.3%

We performed a χ^2 test for Operational Definition I: $\chi^2(2) = 10.6$, $p<005$ (Cramer's coefficient $\phi=0.28$, $p<.01$, N=136). This result indicates that children's performance on Operational Definition I statistically depends on age. The test with Operational Definition II gave the following result: $\chi^2(2)=45.3$, $p<.000$ (Cramer's coefficient $\phi=0.58$, $p<.005$, N=136). The Cramer coefficient indicates that performance on Operational Definition I is only weakly associated with age. In contrast, performance on Operational Definition II and age show a moderate to strong association. The difference between these two Cramer coefficients is statistically significant (t=3.44. p<.001) (Howell, 1997).

Relative Demands of Operational Definitions I and II.

Table 3 shows the number of children that succeeded or failed in either of the two operational definition tasks. Only 9% (N=12) of the children failed both tasks. Forty-seven percent (N=64) of the children succeeded on both tasks. These values testify to the effectiveness of the teaching intervention. Forty-three percent (N=59) of the children succeeded in Operational Definition I and failed in II. In contrast, only 1 child succeeded in Operational Definition II and failed in I. These findings support the sequencing of our curriculum by indicating that Operational Definition II (Lesson 6) is more demanding than Operational Definition I (Lesson 4).

Table 3: Children's performance on Operational Definitions

	Operational Definition I	
Operational Definition II	Failure	Success
Failure	12	59
Success	1	64

To confirm this finding we carried out McNemar's test for the significance of change on the sample of children that were taught all 6 lessons (N=136): $\chi^2(1)=54.2$, p=.000. The result clearly confirms that Operational Definition II is significantly more difficult than Operational Definition I.

Tables 2 and 3 do not include the performance of children in class 5 because this class received the modified intervention (lessons 1-3, 5, 6) and was only tested for Operational Definition II. Class 5 and the group of children listed in Table 2 with age higher than 5 years have very similar average ages. The t-test between these two groups shows that the difference in mean age is not statistically significant. In other words, class 5 is matched to the group of older children in Table 2 in age ($t(38)=-0.6$, $p>.5$). Only 41.3% of the children in class 5 performed successfully in Operational Definition II. In comparison with 85.4% success rate for the older children in Table 2, this is appreciably lower. The children in class 5 performed closer to the 4.5–5 year olds rather than the >5 year olds. The χ^2 test for Operational Definition II ($\chi^2(1)=13.8$, $p<.005$) indicates that there is a statistically significant difference in the performance of class 5 and the older group of children who received the complete intervention. Not teaching lesson 4 and Operational Definition I seems to have influenced these children's performance on Operational Definition II significantly. This would indicate that to some extent Operational Definition I (and lesson 4) functions as a conceptual pre-requisite to successful performance on the Operational Definition II task.

Discussion

It is evident that even though the children across all ages could not initially categorize magnets, the appropriate didactic intervention led them to construct the first operational definition. Our study shows that 79.6% of the first age group, 97.8% of the second age group, and 95.1% of the third group succeeded in constructing the first operational definition.

The nature of the first definition explains the success rate across all ages. Some objects attract others, whereas some other objects do not. Attraction was the only factor to be taken into account. The definition could be constructed by using information from one source only and did not clash with children's conception of magnetism as a substantial property and with their view of causality as an homeopathy. Additionally, it was consistent with the image schema of attracting force and the pulling model.

Only 14.3% of the first group could construct Operational Definition II. In contrast, fifty percent of the second group and 85.4% of the third group succeed in the task. Thus, only children older than 5 years of age succeed. Almost all younger children fail.

In the second definition, some objects sometimes attract and sometimes repel other objects. In this case, there are two factors to be taken into account and the information must be combined for successful categorization. Thus, children who cannot perform logical multiplication are unable to understand it. Magnetism is seen, now, as the result of an interaction between bodies and not necessarily as a property that an object may have by itself. This is a second reason that makes it more difficult for the younger children to succeed. The phenomenon of repulsion further complicates matters. Homeopathy seems to be violated, since the same object can attract and repel other objects. The causal patterns that can explain the phenomenon become more complex.

Children over 5 years of age, having mostly acquired the skill of logical multiplication and being able to coordinate causal schemes, can benefit from instruction and are ready to construct the second definition. We do not claim that the children who succeed, understand the relational nature of magnetism. The fact still remains that a necessary step toward accomplishing this has been undertaken that will allow them to overcome the epistemological obstacle of conceiving all physical properties as substantial properties of matter.

We have elaborated on the crucial issue of scaffolding the learning environment, and argued that successful learning requires that the learner does not process the full complexity of the problem from the beginning. The learning system has the opportunity to learn first the domain's basic features and regularities. These provide the learning system with a code that will allow it to recode the information pertaining to the complex problem.

Children who are cognitively ready (older than 5 years of age) and have been trained with appropriately scaffolded material are expected to exhibit a markedly different performance pattern. To test this, we bypassed lesson four and the first operational definition with a group of 29 students, proceeding directly to the second operational definition. The study shows that only 41,4% succeeded in constructing the second definition, in comparison with 85.4% of the preschoolers of the same age (the third group). Our study, thus, confirms the decisive role of diminishing the cognitive load that the learner initially faces.

References

Barrow, L.H. (1987). Magnet concepts and elementary students' *misconceptions.* In J. Noval (Ed.), *Proceedings of the second international seminar on misconceptions and educational strategies in science and mathematics* (pp. 17-32). Ithaca, NY: Cornell University Press.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: The MIT Press.

Chi, M. T. H. (1992). Conceptual Change within and across Ontological Categories. In R. Giere (Ed.), *Cognitive models of science*. Minnesota University Press.

Clark, A., and Thornton, C. (1997). Trading spaces: Computation, representation, and the limits of uninformed learning. *Behavioral and Brain Sciences*, *20*, 57-66.

Demetriou, A., Efklides, A., & Platsidou, M. (1993). The architecture and dynamics of developing mind: Experiential structuralism as a frame for unifying cognitive developmental theories. *Monographs of the Society for Research in Child Development*, *58* (5-6, Serial No. 234).

Demetriou, A., Efklides, E., Papadaki, M., Papantoniou, A., and Economou, A. (1993a) The structure and development of causal-experiemntal thought. *Developmental Psychology*, *29*, 480-497.

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognitive Science*, *12*, 1-55.

Elman, J. (1991). Learning and development in neural networks: the Importance of Starting Small. *Cognition*, *48*, 71-99.

Erickson, Gaalen (1994). Pupils' understanding of magnetsim in a practical assessment context: The relationship between content, process and progression. In P. J. Fensham, R. F. Gunstone, and R. T. White (Eds.), The content of Science: A constructivist approach to its teaching and learning. The Falmer Press.

Gagliari, L. (1981). Something missing on magnetism? *Science and Children*, *18*, 24-25.

Gentner, D., and Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science*, *10*, 3, 277-300.

Halford, G. S., and Macdonald, C. (1977). Children's pattern construction as a function of age and complexity. *Child Development*, *48*, 1096-1100.

Howell, D. C. (1997) *Statistical Methods for Psychology* (Fourth edition). NY: Duxbury Press pp. 263-265.

Johnson, M. (1987). *The Body in the mind: The bodily basis of meaning, Imagination, and reason*. Chicago, Ill: University of Chicago Press.

Kemler, D. G. (1983). Holistic and analytic modes in perceptual and cognitive development. In Th. J. Tighe and Br. E. Shepp (Eds.), *Perception, cognition, and development: Interactional analyses*. Hillsdale NJ: Lawrence Erlbaum Associates.

McDermott, L. C. and the Physics Education Group (1996). *Physics by iquiry*. N.Y: John Wiley.

Peters. L., Davey, N., Messer, D., and Smith, P. (1999). An investigation into Karmiloff-Smith's RR model: The effects of structured instruction. British Journal of *Developmental Psychology*, *17*, 277-292.

Raftopoulos, A. (1997). Resource limitations in early infancy and its role in successful learning: A connectionist approach. *Human Development*, 40, 5, 293-319.

Selman, R. L., Krupa, M. P., Stone, C. R., and Jacquette, D. S. (1982). Concrete operational thought and the emergence of unseen force in children's theories of electromagnetism and gravity. *Science Education*, *66*:*2*, 181-194.

Shepp, B. E. (1978). From perceived similarity to dimensional structure: A new hypothesis about perspective development. In E. Rosch and B. B. Lloyd (Eds.), *Cognition and categorization.* Hillsdale, NJ: John Willeys and Sons.

Smith, L. B., and Kemler, D. G. (1977). Developmental trends in free classification: Evidence for a new conceptualization of perceptual development. *Journal of Experimental Child Psychology*, *24*, 279-298.

Springer, K., and Keil, F. C. (1991). Early differentiation of causal mechanisms appropriate to Biological and nonbiological kinds. *Child Development*, *62*, 767-781.

Vosniadou, S. (1987). Children and metaphors. *Child Development*, *58*, 870-85.

Testing a computational model of categorisation and category combination: Identifying disease categories and new disease combinations

Fintan Costello (fintan@compapp.dcu.ie)
School of Computer Applications, Dublin City University,
Glasnevin, Dublin 9, Ireland.

Abstract

The diagnostic evidence model gives a computational account of how people classify items in single categories and in category combinations (complex categories formed by combining two or more single categories). This model sets out to explain generativity in category combination (the fact that people can classify items in new category combinations even if they have never seen any examples of those combinations). The model also aims to explain context effects such as overextension in category combination. In an experiment people learned to identify imaginary diseases from artificially-constructed patient descriptions, and then classified new patient descriptions into combinations of those disease categories. The model accurately predicted people's classification scores for patient descriptions in these disease combinations, requiring no free parameters to fit the experimental data. The experiment showed that both generativity and overextension can occur in combinations of artificially-constructed disease categories, and confirmed the model's predictions about when overextension and generativity will occur.

Introduction

The ability to combine mental representations is a basic part of human cognition. For example, to understand a combined phrase such as *pet fish* we must somehow combine our representations of the constituent categories *pet* and *fish*. Category combination is generative: we can understand a new combined phrase such as *pet lobster*, even though we may never have seen an example of a *pet lobster* before. Generativity is important because it allows us to think new thoughts, understand new expressions, and respond to new situations. However, generativity poses a problem for theories of classification in which an item's membership in a category is proportional to its similarity to previously seen exemplars of that category (e.g. the Context theory; Medin & Schaffer, 1978). Since no previously-seen exemplar is available for combined categories such as *pet lobster*, membership in such a category cannot be computed by exemplar similarity (Rips, 1995). While such theories give a good account for classification in single categories, they do not extend well to category combination.

Context effects such as overextension occur reliably in category combination. These effects also pose a problem for current theories of classification. Overextension occurs when people classify an item as a poor member of a constituent category of a combination, but as a good member of the combination as a whole; for example, when people rate goldfish as poor members of the single categories *pet* and *fish*, but as highly typical members of the conjunction *pet fish* (Hampton, 1988). Overextension shows that an item's category membership can change depending on the context in which classification occurs: being poor if the category occurs singly, but good if it occurs as part of a combination. For theories in which classification is based on fixed rules for category membership (e.g. Nosofsky, Palmeri, & McKinley, 1994), these changes in membership are difficult to explain. While such theories apply well to classification in single categories, the do not extend to category combination.

This paper describes a computational model which explains classification in both single and combined categories. This model, called the diagnostic evidence model, also explains generativity and overextension in category combination. This paper also describes an experiment investigating classification, category combination, generativity and overextension in artificial laboratory-learned categories. In this experiment people learned to identify imaginary diseases from artificially-constructed patient descriptions, and then classified new patient descriptions into combinations of those disease categories. Both generativity and overextension occurred reliably in these combinations of artificial categories. The model accurately predicted people's classification scores for patient descriptions in these disease combinations, requiring no free parameters to fit the data. The patterns of overextension and generativity in the experiment closely matched those predicted by the model.

The Diagnostic Evidence Model

The diagnostic evidence model is an extension of a model originally developed to explain how people interpret novel noun-noun phrases (Costello & Keane, 1997, 2000, 2001). The model aims to explain classification in both single categories (see Costello, 2000) and category combinations. The model assumes that people represent categories by storing sets of category members in memory. From these sets, diagnostic attributes for categories are computed: these attributes serve to identify new category members. An item's classification in a single or combined category is a function of the diagnosticity of its attributes for that

category or for the constituent categories of that combination. An item has a high classification score in a category if it has diagnostic attributes of that category. An item has a high score in a combination if it has some attributes diagnostic for one constituent of the combination, and others diagnostic for the other.

Attribute Diagnosticity

Diagnostic attributes are attributes which occur frequently in members of a category, but rarely in that category's contrast set (the set of non-members of that category). These attributes serve to identify members of a category: a new item having an attribute which is diagnostic for a category is likely to be a member of that category. Equation 1 defines the diagnosticity of an attribute x for a category C. Let K be C's contrast set. Let j_X be 1 if an item j has attribute x, and 0 otherwise. $D(x|C|K)$, the diagnosticity of x for C relative to K, is equal to the number of members in C which have attribute x, divided by the total size of C plus the number of items in K which have x:

$$D(x \mid C \mid K) = \frac{\sum_{j \in C} j_x}{|C| + \sum_{j \in K} j_x} \quad (1)$$

If the attribute x occurs in all items in C, but no items in C's contrast set, then x is fully diagnostic for C ($D(x|C|K) = 1$). Such an attribute is a perfect guide to membership of C: a new item having that attribute is most likely a member of C. An attribute which does not occur in all members of C, or which occurs in some members of C's contrast set, will be less diagnostic for the category. Such an attribute will be a poorer guide to membership of C: a new item with that attribute is less certain to be a category member.

Diagnosticity changes in combination

The contrast set is important in computation of attribute diagnosticity: the fewer occurrences of an attribute in the contrast set for a category or combination, higher its diagnosticity will be. The contrast set for a single category consists of all items which are not members of that category. The contrast set for a combined category, however, consists of all items that are not members of any constituent of the combination. This change in contrast set means some attributes that are not diagnostic for a category occurring singly can be diagnostic for that category in a combination. This change in diagnosticity is the basis for overextension in category combination.

Table 1 shows 10 stored members of categories such as *pet* and *fish*, described on 4 dimensions. Computation of attribute diagnosticity can be illustrated using this set of stored category members. Consider the diagnosticity of the attribute <found:house> for the category *fish*. <found:house> occurs in 2 of the 4 members of *fish* in Table 1, and occurs 4 times in the

Table 1. Example items: 10 stored category members.

Item	Categories		Item Attributes			
			FOUND	KEPT	COLOR	PARTS
1		lobster	sea	------	pink	claws
2		lobster	aquarium	tank	pink	claws
3	fish	goldfish	house	tank	gold	scales
4	fish	guppy	house	tank	silver	skin
5	fish	salmon	sea	------	silver	scales
6	fish	shark	sea	------	silver	skin
7	pet	spaniel	house	basket	brown	tail
8	pet	pitbull	house	kennel	black	tail
9	pet	bulldog	house	basket	brown	-----
10	pet	terrapin	house	tank	green	skin

contrast set K_{fish} (the set of items which are not members of the category *fish*). The diagnosticity of <found:house> for *fish* is

$$D(found:house \mid fish \mid K_{fish}) = \frac{2}{4+4} = 0.25 \quad (2)$$

This attribute has a low diagnosticity for the single category *fish*: <found:house> does not identify members of category *fish* well. In the context of the combination *pet fish*, however, the attribute has a higher degree of diagnosticity for *fish*. $K_{petfish}$, the contrast set for the combination *pet fish*, consists of items that are members neither of *pet* nor of *fish* (items 1 and 2 only). <found:house> does not occur in any items in $K_{petfish}$. The diagnosticity of <found:house> for *fish* relative to the contrast set $K_{petfish}$ is thus

$$D(found:house \mid fish \mid K_{petfish}) = \frac{2}{4+0} = 0.5 \quad (3)$$

Attribute <found:house> is thus more diagnostic for *pet fish* than for *fish* alone. Given this, the diagnostic evidence model would predict overextension for the combination *pet fish:* an item such as *goldfish*, which possessed the attribute <found:house>, could be classified as an untypical *fish*, but as a typical *pet fish*.

A logic for evidence

Diagnostic attributes give evidence for an item's classification in a category. Items usually contain a number of different attributes, however, which may be more or less diagnostic for the category in question, or diagnostic for other categories. The diagnostic evidence model uses a continuous-valued logic to combine the diagnosticity of multiple attributes. This logic assumes continuous variables with values between 0 and 1, and uses the logical operations

$$NOT\ A = 1 - A \quad (4)$$

$$A\ AND\ B = AB \quad (5)$$

$$A\ OR\ B = 1-(1-A)(1-B) \quad (6)$$

These equations can be justified by considering the operations *AND*, *OR*, and *NOT* for samples of independent variables. Suppose A is true in 75% of samples, and B

is true in 50% of samples. Then the probability of *NOT A* being true is *0.25 (1-0.75)*. The probability of *A AND B* being true is *0.375 (0.75 x 0.5)*: *A* is true in 75% of samples, *B* is true in 50% of those. Finally, the probability of *A OR B* being true is *0.875 (1-(1-0.75) x (1-0.5))*: *A* is false in 25% of samples, *B* is false in 50% of those, and thus *A OR B* is true in 87.5% of samples.

Combining attribute diagnosticities

To compute an item's overall evidence for membership in a category, the diagnosticity of the item's attribute are combined using the equation for *OR* (Equation 6). An item *i* with a set of attributes x_1, x_2, x_3 will be a member of category *C* if x_1 *OR* x_2 *OR* x_3 is diagnostic for C. This is formalised in Equation 7. Let *A* be the set of attributes of item *i* and let *D(x|C|K)* be the diagnosticity of attribute *x* for *C*. Then *E(i|C|K)*, the overall evidence for classifying item *i* as a member of *C*, is

$$E(i \mid C \mid K) = 1 - \prod_{x \in A} (1 - D(x \mid C \mid K)) \qquad (7)$$

If an attribute *x* strictly defines a category *C* (occurs in all members of *C* and never occurs outside *C*), then *x* is perfectly diagnostic of *C (D(x|C|K) = 1)*. If any item *i* possesses attribute *x* , then by Equation 7 *E(i|C|K)* will be 1, and the item *i* will definitely be a member of category *C*. In categories which have no single defining attribute but rather a range of attributes of medium diagnosticity, Equation 7 combines evidence from different attributes in computing evidence for category membership: the more diagnostic attributes the item has, the higher its degree of membership will be. This fits with the observed family resemblance structure of natural categories (Rosch, 1978). The relationship between diagnosticity and membership is supported by Rosch & Mervis' (1975) finding that people's judgements of an item's typicality in a category rises with the number of the item's diagnostic attributes.

Diagnostic evidence in combinations

In the diagnostic evidence model, an item will be a member of a combined category if it gives evidence for membership in each constituent category in that combination: if it has some attributes diagnostic for one constituent of the category, and other attributes diagnostic for the other. In computing an item's membership in a combined category, the model uses the equation for *AND* to combine the item's evidence for membership in each constituent. An item *i* will be classified as member of a combined category $C_1 \ldots C_N$ if it gives evidence for membership in C_1 *AND* evidence for membership in C_2 *AND* evidence for membership in C_3 and so on. Formally, *E(i|C1...CN|K1...N),* the evidence for classifying *i* as a member of $C_1 \ldots C_N$, is

$$E(i \mid C_1 \ldots C_N \mid K_{1\ldots N}) = \prod_{n=1\ldots N} E(i \mid C_n \mid K_{1\ldots N}) \qquad (8)$$

Table 2. Classification of the item *goldfish* in single categories *pet* and *fish* and combination *pet fish.*

Evidence for membership in		Attribute Diagnosticity			
		FOUND house	KEPT tank	COLOR golden	PART scales
pet singly :	**0.7**	0.7	0.1	0.0	0.0
fish singly:	**0.8**	0.2	0.3	0.2	0.5
***pet fish*:**					
constituent *pet*	**1**	1.0	0.2	0.0	0.0
constituent *fish*	**0.9**	0.5	0.4	0.2	0.5
Pet fish overall:	**0.9**				

where the contrast set $K_{1\ldots N}$ is the set of items not in any category $C_1 \ldots C_N$. In this equation an item *i* gives evidence for membership in each constituent of a combination if it has attributes diagnostic for each. Note that, in computing the evidence for membership in each constituent category (r.h.s. in Equation 8), the contrast set for the combination as a whole is used. In computing membership in those categories occurring singly, their single contrast sets would be used.

Table 2 illustrates the diagnostic evidence model by showing the computed membership for the item *goldfish*, which has attributes <found:house>, <kept:tank> <color:golden> and <part:scales>, in the single categories *pet* and *fish* and in the combination *pet fish*. The diagnosticity of the item's attributes for single categories and for constituents of the combination are listed in columns under those attributes. The item's membership scores in the single categories and the constituents of the combination are computed from those diagnosticities (shown in bold, to the left of those diagnosticities). At the bottom of Table 2 is the item's overall membership score in the combination (obtained by multiplying its constituent membership scores).

Explaining overextension and generativity

In the diagnostic evidence model, overextension arises if some attributes have low diagnosticity for a single category but high diagnosticity for that category in a combination. Table 2 illustrates this overextension. The item in Table 2 has a higher overall membership score in the combination *pet fish* than in the categories *pet* or *fish* presented singly, because the item's attributes are more diagnostic for the combination than for the single categories. For example, <found:house> has lower diagnosticity for the single category *fish*, but higher diagnosticity in context of the combination *pet fish* (it occurred often in the contrast set for the single category *fish*, but not in the contrast set for *pet fish*). The model thus predicts overextension for that item.

The diagnostic evidence model gives a generative account of category combination, in which an item can be classified in a new combination even if no previous examples of that combination have been seen. An item

Table 3. Training materials for learning diseases.

Training Item	Item features D1	D2	D3	Member of Category or Combination
1	A	X	C	*A*
2	A	Y	Y	*A*
3	A	A	X	*A*
4	Y	A	Y	*A*
5	X	A	B	*A&B*
6	A	B	X	*A&B*
7	Z	B	B	*B*
8	X	B	B	*B*
9	Y	X	B	*B*
10	Z	Y	B	*B*
11	C	A	Y	*C*
12	C	X	B	*C*
13	C	Y	C	*C*
14	C	A	C	*C*
15	C	X	C	*C*
16	X	Y	C	*C*

is classified in such a combination if it has diagnostic attributes for each constituent category in the combination: some attributes diagnostic for one constituent, other attributes diagnostic for the other. For example, in Table 1, there are no stored members of the combination *pet lobster*. However, an item could be classified as a good member of the combination *pet lobster* if it possessed the attributes <has-part:claws> (diagnostic for *lobster* in Table 1) and <found:house> (diagnostic for *pet*). The model thus predicts generativity for that combination. The next section describes an experiment testing the diagnostic evidence model's predictions about classification, overextension, and generativity in category combination. This experiment uses artificial categories in the domain of disease diagnosis.

Disease Diagnosis: An Experiment

Most experiments investigating category combination examine how natural-language categories are combined. The current experiment examines category combination with artificial, laboratory-generated categories representing imaginary diseases. In this experiment, every subject was given a set of 16 patient descriptions (16 training items), each with 3 symptoms and each having a given disease or disease combination. The abstract distribution of symptoms in training items was identical for all subjects, and is shown in Table 3. The training materials used ill-defined categories: no symptom perfectly indicated any disease. In the training phase of the experiment, subjects used these training items to learn to identify diseases. In the transfer phase subjects were given new patient descriptions (transfer items) and asked, for each disease and each possible disease combination, to indicate whether the new item was a member of that disease category or disease combination.

Method

Subjects. 19 Dublin City University undergraduates.
Materials. Each subject received a set of 16 patient-description cards (training items) with the abstract structure shown in Table 3. In these, abstract attribute A (on any dimension) is most diagnostic for category *A*, attribute B for category *B*, and C for category *C*. Each subject received a different set of patient descriptions, generated via a unique mapping from abstract attributes to concrete symptoms. For example, for one subject, attribute <A> on dimension D1 became symptom *eyes:puffy*; <A> on D2 became *skin:flaking*, and <A> on D3 became *mucles:taut*. For other subjects the attributes were mapped to different symptoms.

Procedure. In the training phrase subjects spent 15 minutes learning to identify diseases by studying their 16 patient-description cards. Subjects were then shown, in random order, patient descriptions with the same symptoms as those they had learned, but sometimes with incorrect diagnoses. Subjects indicated whether diagnoses were correct and incorrect. If a subject got a diagnosis wrong, they were shown the correct answer. The transfer phase of the experiment examined subjects classification of 5 new patient descriptions (the transfer items). Table 4 gives abstract representations for these items. Each subject's transfer items were formed by applying their attribute-to-symptom mapping to this representation. Each item was presented 6 times, each time with a different single or combined category. Subjects rated the given item as a member or non-member of the given category or combination, using a –10 to +10 rating scale, with a positive rating indicating membership and a negative rating non-membership.

Results

Analysis of subject's performance in the training phase showed that most had no problem in classifying items. One subject got most of the training-phase test classifications wrong and was excluded from analysis.
The 2nd-last column in Table 4 ('classification probability: observed') shows the observed probability (proportion) of subjects rating each transfer item as a member of the given combination. (For space reasons the corresponding data for single categories are not shown.) For example, the observed probability of transfer item <ABY> being classified in combination *A&B* was 0.5: 50% of subjects rated that item as a member of that combination.
The data in Table 4 shows that subjects responded consistently to items. For example, there were some items which had high observed classification probabilities for particular combinations, indicating that many subjects agreed that those items did belong in

Table 4. Observed and predicted classification of the 5 transfer items in 3 different category combinations.

Item			Combination	Classification probability	
D1	D2	D3		Observed	Predicted
A	B	Y	A&B	0.50	0.47
A	B	Y	A&C	0.11	0.13
A	B	Y	B&C	0.11	0.07
C	Y	B	A&B	0.06	0.19
C	Y	B	A&C	0.28	0.21
C	Y	B	B&C	0.72	0.77
Y	A	C	A&B	0.22	0.14
Y	A	C	A&C	0.50	0.50
Y	A	C	B&C	0.17	0.17
X	B	C	A&B	0.28	0.23
X	B	C	A&C	0.17	0.27
X	B	C	B&C	0.39	0.42
X	X	B	A&B	0.28	0.29
X	X	B	A&C	0.11	0.15
X	X	B	B&C	0.39	0.45

those combinations. For example, 72% (0.72 observed classification probability) of participants identified item <CYB> as a member of combination *B&C*: most subjects agreed in classifying that item as a member of that combination. Conversely, a number of items had very low observed classification probabilities for particular combinations, indicating that a large proportion of subjects agreed that those items did not belong in those combinations. For example, only 6% (0.06 observed classification probability) of participants classified item <CYB> as a member of combination *A&B*. The remaining 94% of participants indicated that the item did not belong in that combination. Because each subject's patient descriptions used a unique mapping from abstract attributes to symptoms, this consistency depended only on the distribution of those symptoms in the learned categories.

Model fit. To apply the diagnostic evidence model to the experimental materials, the equations described earlier were used to compute the classification score for each of the transfer items in Table 4 in every possible single and combined disease category. The diagnosticity of each item's attributes for each category was computed from the distribution of those attributes in the training items shown in Table 3. The last column in Table 4 ('classification probability: predicted') shows computed classification scores for each item in each combined category. These computed scores were compared with the observed probability with which people classified the items as members of each combination. There was a strong correlation between the predicted and observed classification scores for the combined categories (r=.95, p <.01, *%var*=.9). Comparing predicted and observed classification scores for items across all single and all combined categories also showed a significant correlation (r=.85, p <.01, *%var*=.73). No free parameters were used to fit the model's classification scores to the experimental data.

Table 5. The 6 item-combination-constituent triplets for which overextension occurred or was predicted.

Item	Combination	Constituent	Overextension	
			Observed	Predicted
A B Y	*A&C*	*C*	Yes (0.50)	Yes
C Y B	*A&C*	*A*	Yes (0.56)	Yes
C Y B	*B&C*	*B*	Yes (0.67)	Yes
Y A C	*A&C*	*C*	Yes (0.56)	**No**
X X B	*A&B*	*A*	**No** (0.41)	Yes
X X B	*B&C*	*C*	Yes (0.61)	Yes

Generativity. The generativity of category combination was examined by comparing the classification of transfer items in the combinations *A&B*, *A&C*, and *B&C*. In the training phase, subjects saw examples of the combination *A&B* but not of the other combinations. If combination is not generative, participants will only be able to identify items as members of the previously-seen combination *A&B*, but not as members of the other two combinations. Table 3 shows that 72% of participants classified the item <CYB> as a member of the previously-unseen combination *B&C*. More participants classified the item in that new combination than classified the item <ABX> in the previously-seen combination *A&B*. There was no significant difference between the number of items classified in the previously-seen combination *A&B* and in the other, new combinations. This supports a generative view of category combination.

Overextension. The occurrence of overextension in the experimental data was analysed at the individual subject level. Overextension was taken to have occurred every time an individual subject gave a particular transfer item a higher score as a member of a combined category than they gave that item as a member of one of the constituent categories of that combination. For example, if a given subject gave the transfer item <ABY> a high classification score as a member of *A&C*, but the same subject gave <ABY> a lower classification score as a member of *C*, that would be taken as a case of overextension (an overextension response). In the experiment there were 5 different transfer items, each of which was classified in 3 different category combinations (*A&B*, *A&C*, *B&C*), where each combination had two different constituent categories. There were thus 30 possible cases in which overextension could arise. Out of those 30 cases, there were 5 in which 50% or more subjects produced overextension responses. Table 5 shows all item-combination-constituent triplets for which overextension

either occurred in the experiment or was predicted by the model. The 5 cases with at least 50% overextension responses in the experiment are indicated in Table 5 by a "Yes" in the 2nd-last column (with the proportion of overextension responses for each case). For these cases, at least 50% of subjects rated the given item as a better member of the given combination than of the given constituent category presented singly. In a significant number of cases subjects rated the given item as a member of the combination, but as a non-member of the constituent category. For example, 44% of subjects rated the item <CYB> as a member of the combination *B&C*, but the same subjects rated that item as a non-member of the constituent category *B*. These cases show that overextension occurs reliably even for artificial categories. (Order of presentation had no reliable influence on overextension in these cases.)

To analyse the model's predictions about overextension, the model's computed classification scores for all transfer items in all category combinations were compared with the scores for those items in the constituents presented singly. If an item had a higher classification score in a combination than in one of its constituent categories presented singly, the model predicted overextension for that item. Again, there were 30 cases in which overextension could happen: out of those 30 cases, the model predicted overextension in only 5 cases. These are indicated by a "Yes" in the last column in Table 5. Of the 5 cases in which overextension occurred in the experiment, 4 were cases in which overextension was predicted by the model. Of the remaining 25 cases in which overextension was not observed, 24 were cases in which the model predicted overextension would *not* occur. The model accurately predicted the occurrence and non-occurrence of overextension in these materials.

Discussion and Conclusions

The results obtained in the above experiment are important for a number of reasons. They show that both overextension and generativity occur even for combinations of artificial laboratory-learned categories. Previous research has investigated these factors in natural-language category combinations alone. They show that the patterns of overextension and generativity seen in the experiment have a close quantitative match with those predicted by the diagnostic evidence model. Other models give a looser qualitative account of overextension and generativity. Finally, these results show that the diagnostic evidence model accurately predicts people's classification of items in both single categories and category combinations, needing no free parameters to fit the data. Other models typically apply either to single categories or to category combination, but not to both. These models typically require a number of free parameters to fit the relevant data.

The diagnostic evidence model, in accounting accurately for the results of the above experiment, represents an advance on other current theories of classification. However, there are results which the model cannot currently explain. For example, studies show that people can learn correlations between pairs of attributes and use those correlations in classification (Medin, Altom, Edelson, & Freko, 1982). The diagnostic evidence model, because it has no mechanism for learning correlations between attributes, cannot account for these results. In future work the model will be extended to learn attribute correlations by forming new "composite" attributes, and to use those attributes in classification. This may allow the model account for these findings.

References

Costello, F. J. (2000). An exemplar model of classification in single and combined categories. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society.*

Costello, F. J., & Keane, M. T. (1997). Polysemy in conceptual combination: Testing the constraint theory of combination. In *Proceedings of the nineteenth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Costello, F. J., & Keane, M. T. (2000). Efficient creativity: Constraint guided conceptual combination. *Cognitive Science,24(2).*

Costello, F. J., & Keane, M. T. (2001). Testing two theories of conceptual combination: Alignment versus diagnosticity in the comprehension and production of combined concepts. J*ournal of Experimental Psychology: Learning, Memory & Cognition, 27(1),* 255-271.

Hampton, J. A. (1988). Overextension of conjunctive concepts: Evidence for a unitary model of concept typicality and class inclusion. *Journal of Experimental Psychology: Learning, Memory and Cognition, 15,* 55-71.

Medin, D. L., Altom, M. W., Edelson, S. M., & Freko, D. (1982). Correlated symptoms and simulated medical classification. *Journal of Experimental Psychology: Learning, Memory and Cognition, 8,* 37-50.

Medin, D. L. & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85(*3), 207-238.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. K. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101*, 53-79.

Rips, L. J. (1995). The current status of research on concept combination. *Mind & Language, 10*, 72-104.

Rosch, E. & Mervis, C. D. (1975). Family resemblance studies in the internal structure of categories. *Cognitive Psychology, 7*, 573-605.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.) *Cognition and Categorization.* Hillsdale, NJ: Erlbaum.

Exploring Neuronal Plasticity: Language Development in Pediatric Hemispherectomies

Stella de Bode (sdebode@ucla.edu)
UCLA, Department of Linguistics; 405 Hilgard Ave
Los Angeles, CA 90095

Susan Curtiss (scurtiss@ucla.edu)
UCLA, Department of Linguistics; 405 Hilgard Ave
Los Angeles, CA 90095

Abstract

We investigated the categories of neural plasticity and the genesis of the neural representation for language in population of 43 pediatric hemispherectomies. We have chosen to correlate language outcomes with the stages of neuronal plasticity rather than age at insult because of the unavoidable confound between the latter and etiology (Curtiss and de Bode, submitted). We argue that by examining the neural substrate for language and language outcomes post-hemispherectomy, it is possible a) to investigate the progression of neural representation from pluripotential and distributed to localized and specialized and b) to accurately predict language outcomes.

Introduction and Rationale

It is still unclear whether neural systems underlying adult organization for language crucially differ from their respective counterparts in the young brain. Though the assumption of complete and rapid recovery of children after brain lesions has been abandoned by the majority of researchers, there is no question that the rate and extent of reorganization in children differ from adults recovering from similar insults. The two most obvious hypotheses explaining this phenomenon make two different sets of assumptions. First, it is possible that language representation in a young brain is not identical to its adult counterpart. Indeed, more diffuse brain organization of the immature brain is suggested both by recent brain imaging studies and language acquisition research in clinical and normal populations (Dapretto, Woods, & Bookheimer, 2000; Mills, Coffey-Corina, & Neville, 1993; Papanicolaou, DiScenna, Gillespie, & Aram, 1990). Under this hypothesis faster recovery rates in children may be explained by the fact that functional localization and cortical commitment have not yet reached their peak, i.e. their adult pattern. An alternative explanation does not need to assume brain organization that is different from adults. Empirical support for this hypothesis is provided by investigations of childhood acquired aphasia. This research indicates the presence of adult-like neural representation for language and similar consequencies of brain damage in children and adults (Paquier & Van Dongen, 1998). Thus it is possible that more efficient reorganization is achieved due to neural plasticity of a young brain, in other words, with the help of the same mechanisms that are already in place guiding and supporting brain maturation in the first decade of life.

The two accounts need not be mutually exclusive. It is possible that what seems like wider functional distribution is, in fact, the reflection of both exuberant neuronal connectivity and increased neuronal excitation characteristic of an immature brain. This suggestion is supported by the findings of some recent brain imaging studies. Dapretto et al. (2000) demonstrated that both phonological and semantic conditions activated similar though not completely identical areas in adults and children. Furthermore, cortical areas activated only by specific linguistic tasks in adults showed reliable activation during all tasks in children. The authors interpret these findings in terms of increased functional specialization with development and redundancy in the neural system subserving language early in development. Taking these conclusions one step further, we suggest that the dichotomy of 'pluropotential and distributed' versus 'specialized and localized' exists only on the functional level. On the neurobiological level, language representation in children is similar to adults, but this similarity is masked by diffuse connectivity and exuberant synaptic proliferation that characterize the young brain.

For the purpose of this paper we assume that an innate endowment and cortical representation for language are present from birth. We also assume that quantitative differences of an immature cortex lead to some qualitative differences (such as pluripotential

cortex and distributed functional organization in infants) but represent a developmental continuum within the framework of similar language representation in children and adults. What do we attribute to the processes underlying quantitative differences between the young and mature brains? Similar to animal research, morphometric and brain imaging studies (EEG, glucode metabolism, blood flow volumes, etc.) in humans imply the presence of the period of massive overproduction of synapses, dendritic arbors and exuberant connectivity. This period, known as the Critical Maturation Period, leads to the next stage of development - the process of elimination when neuronal/synaptic numbers, density, connectivity are adjusted to their respective adult values. Though there is no complete data regarding the exact timetables of these events for the entire brain, it is known, for example, that these overproduction/adjustment processes in the frontal lobes continue into adolescence (Huttenlocher, 1993). The outline of our hypothesis is shown in Table. 1:

Table 1. Rationale for our hypothesis

	Young brain	Adult brain
Neurobiological level	Similar language representation	
	Morphological/Quantitative changes underlying brain maturation (synaptogenesis, dendritic proliferation, neuronal volume adjustment)	
Functional level	Pluripotential & distributed	Specialized & localized

Methods

Subjects consisted of 43 patients who underwent hemispherectomy for intractable seizures at the UCLA Medical Center. Etiology was catalogued according to the following breakdown: developmental pathology - 28 subjects (hemimegalencephaly - HM, cortical dysplasia/multilobar involvement - ML, and prenatal infarct); acquired pathology - 15 subjects (Rasmussen's encephalitis - RE and postnatal infarct). Postoperative spoken language outcome was rated based on spontaneous speech samples from 0 = no language to 6 = fluent mature grammar. Language scores were defined on the basis of stages in normal language development. The complete information regarding the breakdown of our population is shown in Table 3.

Discussion

Based on the animal studies we suggest that the Critical Maturation Period in humans is limited by the following thresholds: the lower threshold that is characterized by the completion of neuro/morthogenesis and establishment of experience independent connectivity; and the upper threshold of the completion of the period of neuronal/synaptic adjustment. Next, following Greenough et al. (1999) we assume that the following components underlie functional and neurological maturation of language: (1) developmental processes that are insensitive to experience, i.e. the genetic envelope of predetermined plasticity; (2) an experience-expectant period of neuronal plasticity also known as the Critical Maturation Period; and (3) an experience-dependent period of neuronal plasticity which underlies the ability to encode new experiences throughout the lifespan (Table 2). We thus hypothesized that superimposing effects of specific etiologies on these developmental stages would allow for more accurate prediction of language outcomes following hemispherectomy, since in our model functional reorganization reflects underlying neurobiological reorganization.

Our results confirmed our hypothesis in that postoperative language outcomes correlated with etiology. This would be expected since as shown in Table 2 different etiologies result in different potential for recovery (due to timing and extent). Developmental plasticity, i.e. reinnervation and neuronal sparing, seem to be more efficient in etiologies with later onset. In addition, when pathology disrupts genetically determined processes (as in hemimegalencephaly and cortical dysplasia) functional development seems to be particularly compromised. Thus the best language scores were found in Rasmussen's encephalitis and the poorest in hemimegalencephaly. Moreover, etiology

(developmental or acquired) consistently emerged as a significant variable distinguishing linguistic outcomes in all statistical analyses. In all cases it was possible to predict post-surgery language outcomes by considering the effect of specific etiologies within the framework of the categories of neural plasticity. It should be noted that we have deliberately chosen to relate functional outcomes and the broad categories/stages of neuronal plasticity instead of providing direct correlations with age at insult. It is our belief that in such correlations the confound between etiology and age at insult is unavoidable (Curtiss, de Bode and Mathern, submitted).

The rate and quality of neuronal reorganization reflected in language outcome also confirmed the left hemisphere's predisposition to support language, since children with an isolated right hemisphere had significantly more problems acquiring/restoring their language. Importantly, however, though age at surgery for two of our RE children was as old as 12, neither of them has remained mute after left hemispherectomy, suggesting that language specialization had not yet reached its peak, and reorganization was still possible. Our preliminary research also indicates that even in the most severely compromised cases, language development follows the normal course of language acquisition albeit on a prolonged scale. These findings lead us to suggest that innate language universals are resilient to brain damage, although language representation in the brain does not seem to be anatomically-bound to the left hemisphere only.

Table 2. The impact of specific etiologies on the categories of neural plasticity

Stages/Etiology	Genetic Envelope (innate constraints specifying cortex differentiation including ensembles that would support language-related properties)	Experience-Expectant Period (=Critical Maturation Period, input-dependent period of maximum plasticity)	Experience-Dependent Period (plasticity underlying the ability to incorporate new experiences throughout the lifespan)
Normals	normal	birth – 12 years, reduced vulnerability to injury	normal, life-long
Hemimegalencephaly	affected	increased vulnerability	Limited in most cases, thus lowered FSIQ
Cortical Dysplasia	affected-to-normal	variable	
Infarct prenatal	affected-to-normal	variable	
Infarct postnatal	normal	reduced vulnerability to injury similar to normals	
Rasmussen's Encephalitis	normal	reduced vulnerability to injury similar to normals	

Acknowledgements

We are grateful to all the children and their parents who have graciously agreed to participate in this study.

References

Curtiss, S., de Bode, S., and Mathern, G.W. (2000). Spoken language outcomes after hemispherectomy: factoring in etiology. Brain and Language, submitted.

Dapretto, M., Woods, R.P., and Bookheimer, S.Y. (2000). Enhanced cortical plasticity early in development: Insights from an fMRI study of language processing in children and adults. Paper presented at the Annual Meeting of Neuroscience Society, Los Angeles.

Greenough, W.T., Black, J.E., Klintsova, A., Bates, K.E., and Weiler, I.J. (1999). Experience and plasticity in brain structure: possible implications of basic research findings for developmental disorders. In S.H. Broman & J.M. Fletcher (Eds.), The Changing Nervous System (pp. 57-72). New York: Oxford University Press.

Huttenlocher, P.R. (1993). Morphometric study of human cerenral cortex development. In M.H. Johnson (Ed.), Brain Development and Cognition (112-124). Oxford: Blackwell.

Mills, D. L., Coffey-Corina, S. A., and Neville, H. J. (1993). Language acquisition and cerebral specialization in 20-months-old infants. Journal of Cognitive Neuroscience, 5 (3), 317-334.

Papanicolaou, A., DiScenna, A., Gillespie, L., and Aram, D. (1990). Probe-evoked potential findings following unilateral left-hemisphere lesions in children. Archives of Neurology, 47, 562-566.

Paquier, P. F. and Van Dongen, H. R. (1998). Is acquired childhood aphasia atypical? In P. Coppens, Y. Lebrun, & A. Basso (Eds.), Aphasia in Atypical Populations (pp. 67-117). New Jersey: Lawrence Erlbaum.

Table 3. Subjects

No/Sex	Side 1-L, 2-R	Post-op (years)	Age/onset (years)	Age/surgery (years)	Sz control 1yes 2no	SLR 1to 6
			Hemimegalencephaly			
1M	1	5.2	0.05	2.8	1	4
2M	1	10.1	0.08	3.3	2	0
3M	1	7.8	0.01	0.25	1	0
4F	2	3.3	0.5	2.6	2	1
5F	2	4.3	0.02	2.1	2	0
6F	2	6.7	0.01	0.41	2	1
7M	2	6.2	0.01	1.5	2	0
			Cortical Dysplasia/Multilobar Involvement			
8M	1	3.1	0.5	1.6	1	3
9F	1	5.1	0.01	1.4	2	6
10M	1	8.0	0.01	0.7	1	6
11M	1	9.3	0.01	1.4	2	2
12M	1	5.8	0.01	1	1	4
13M	1	7.2	0.5	1.5	1	6
14F	1	7.4	0.05	0.4	2	5
15M	1	8.1	0.1	0.75	1	3
16F	2	5.6	0.01	0.3	2	3
17F	2	5.3	0.4	0.75	1	5
18F	2	6.1	0.01	1.1	2	2.5
19M	2	8.6	0.75	3.8	2	1
			Rasmussen's Encephalitis			
20M	1	4.7	3.3	4.58	2	4
21M	1	4.3	2.25	3.5	1	4
22M	1	4.2	2.9	5.95	2	3
23M	1	4.11	10.3	12.75	1	5
24F	1	2.0	5	10	1	5
25F	1	3.1	5.5	6.91	2	5
26F	2	8.7	4.75	5.7	1	6
27F	2	12.1	4.18	14	2	6
28F	2	5.9	11	17.3	2	6
29M	2	5.1	2.05	3.41	1	5.5
			Infarct			
30F	1pre-natal	0.6	0.01	6.9	1	4
31M	1post	5.1	3	9.5	2	4
32M	1post	10.2	0.8	6.2	1	5
33M	1pre	4.11	0.25	2.6	1	3
34F	1pre	3.1	0.02	1.3	2	0
35M	1pre	7.8	0.6	8.6	2	0
36F	1pre	8.9	0.01	4	1	5
37M	1pre	5.2	0.5	9.75	1	6
38F	1post	8.8	1.5	6.75	1	5
39F	2pre	8.1	0.3	0.8	2	0
40M	2post	4.9	4	7.75	2	5
41M	2post	11.2	0.6	2.2	1	6
42F	2pre	7.9	1.2	4.25	1	4
43F	2pre	8.1	0.16	5.1	2	5

‘Does pure water boil, when it’s heated to 100°C?’: The Associative Strength of Disabling Conditions in Conditional Reasoning

Wim De Neys (Wim.Deneys@psy.kuleuven.ac.be)
Department of Psychology, K.U.Leuven, Tiensestraat 102
B-3000 Leuven, Belgium

Walter Schaeken (Walter.Schaeken@psy.kuleuven.ac.be)
Department of Psychology, K.U.Leuven, Tiensestraat 102
B-3000 Leuven, Belgium

Géry d’Ydewalle (Géry.dYdewalle@psy.kuleuven.ac.be)
Department of Psychology, K.U.Leuven, Tiensestraat 102
B-3000 Leuven, Belgium

Abstract

Reasoning with conditionals involving causal content is known to be affected by retrieval of alternative and disabling conditions. Recent evidence indicates that besides the number of stored conditions, the relative strength of association of the alternative conditions with the consequent term is another important factor that affects the retrieval process. In this study we examined the effect of the strength of association for the disabling conditions. We identified causal conditionals for which there exists only one highly associated disabler. With these conditionals we constructed conditional inference problems in which the minor premise was expanded with the negation of a strongly or weakly associated disabler. Results indicate that strength of association of the disabling conditions is affecting reasoning performance: Acceptance of Modus Tollens increased when there was no strongly associated disabler available.

Introduction

Conditional reasoning has attracted a lot of interest from cognitive scientists studying human reasoning. Conditional reasoning consists in making inferences on the basis of ‘if p then q’ sentences. In a conditional inference task people are usually asked to assess four kinds of arguments: Modus Ponens (MP, ‘if p then q, p therefore q’), Modus Tollens (MT, ‘if p then q, not q therefore not p’), Denial of the Antecedent (DA, ‘if p then q, not p therefore not q’), and Affirmation of the Consequent (AC, ‘if p then q, q therefore p’).

Under the material implication interpretation of standard logic, MP and MT are considered valid inferences while DA and AC are regarded as fallacies. Much of the work on conditional reasoning has tried to identify the factors that influence performance on these four problems (for a review, see Evans, Newstead, & Byrne, 1993).

A growing body of evidence is showing that peoples knowledge about the relation between the p (antecedent) and q (consequent) part of the conditional has a considerable effect on the underlying reasoning process (e.g., Byrne, Espino, & Santamaria, 1999; Markovits, 1984; Newstead, Ellis, Evans, & Dennis, 1997; Rumain, Connell, & Braine, 1983; Thompson, 1994).

In the case of reasoning with conditionals involving causal content (e.g., ‘If cause p, than effect q’) seminal work has been done by Cummins and her colleagues (1995; Cummins et al., 1991). Following Byrne (1989), Cummins examined the effect of the alternative and disabling conditions of a causal conditional. An alternative condition is a possible cause that can produce the effect mentioned in the conditional while a disabling condition prevents the effect from occurring despite the presence of the cause. Consider the following conditional:

If the brake is depressed, then the car slows down

Possible alternative conditions for this conditional are:

running out of gas, having a flat tire, shifting the gear down…

The occurrence of these conditions will result in the car slowing down. The alternatives make it clear that it is not necessary to depress the brake in order to slow the car down. Other causes are also possible.

Possible disabling conditions are:

a broken brake, accelerating at the same time, skid due to road conditions…

If such disablers are present, depressing the brake will not result in the slowing down of the car. The disablers make it clear that it is not sufficient to depress the brake

in order to slow down the car. There are additional conditions that have to be fulfilled.

When people (fallaciously) accept DA and AC inferences, they fail to see that there are other causes that may lead to the occurrence of the effect beside the original stated one. Cummins (1995) and Cummins et al. (1991) found that peoples acceptance of DA and AC inferences decreased for conditionals with a high number of possible alternative conditions. This showed that a crucial factor in making the fallacious inferences is the number of alternative causes people can think of. In addition, she found that the number of disabling conditions affected the acceptance of the valid MP and MT inferences: If there were many conditions that could disable the relation between antecedent and consequent, people tended also to reject the valid inferences.

Recently, Quinn and Markovits (1998) have identified another factor that may influence reasoning with causal conditionals. They showed that not only the number of alternative conditions is important, but also what they call the 'strength of association' of the alternative conditions. Quinn and Markovits developed a framework (see also Markovits, Fleury, Quinn, & Venet, 1998) where reasoning performance is being linked to the structure of semantic memory. In this framework it is assumed that, when confronted with a causal 'if p then q' conditional, reasoners will access a causal structure in semantic memory that corresponds to 'ways of making q happen' (i.e., alternative conditions). Within the structure, there will be causes that are more strongly associated with q than others. The more strongly associated a specific cause is, the higher the probability that it will be retrieved by the semantic search process.

Quinn and Markovits (1998) measured strength of association by frequency of generation: In a pretest, participants were asked to write down as many potential causes for a certain causal consequent (effect, e.g., 'a dog scratches constantly'). This allowed the construction of conditionals with a strongly (e.g., 'If a dog has fleas, then it will scratch constantly') and weakly (e.g., 'If a dog has skin disease, then it will scratch constantly') associated cause. With the weak conditional, reasoners will be able to activate the strongly associated cause, while they will have to activate some other, less closely associated term for the strong conditional. Thus, it will be more difficult to retrieve an alternative condition in case of the strong conditional, which would lead to a greater acceptance of DA and AC inferences. The results of the study were consistent with the predicted response pattern.

The identification of the strength of association effect raises the question whether this effect is also present for the disabling conditions. Indeed, although knowledge of disabling conditions is also stored in semantic memory, Quinn and Markovits (1998) restricted their case to an analysis of the alternative conditions. Cummins (1995) already showed that both the number of alternatives and disablers is affecting reasoning performance. In addition, Elio (1998) has shown that the process of disabler retrieval is not only important in conditional reasoning but also in the field of belief revision and non-monotonic reasoning: Belief in a conditional after contradiction was lower when people could find many disablers. Thus, both for reasoning psychologists and the psychological and AI community studying belief revision, examining the effect of associative strength of disablers can identify a new factor affecting the crucial disabler retrieval. Therefore, we examined in this study whether Quinn and Markovits' strength of association effect also generalized to the disabling conditions.

The framework developed by Quinn and Markovits (1998) was adopted and extended to the disabling conditions. It was assumed that when presented a causal conditional, people will not only access a causal structure with alternative conditions but also one that corresponds to 'ways that prevent q to occur' (see Vadeboncoeur & Markovits, 1999). When such disabling conditions are retrieved, p will no longer be perceived as a sufficient condition for q what renders the MP and MT conclusions uncertain.

In a generation task we identified strongly and weakly associated disablers for a number of conditionals. We constructed experimental items by expanding the original antecedents of the conditionals with the negation of the strongly or weakly associated disabler. Suppose that for a certain conditional we find that S is a strongly associated disabler, while W is a weak one. This allows the construction of the expanded conditionals: 'If P and not S, then Q' (strongly expanded) and 'If P and not W, then Q' (weakly expanded). These expanded conditionals have an equal number of possible disablers (i.e., the original number minus one). However, reasoners presented with 'If P and not W, then Q' will still be able to activate the strongly associated disabler S, while with 'If P and not S, then Q' they will have to activate a less closely associated one. Thus, it will be harder to access and retrieve disablers for the strong conditionals. This access-to-disablers manipulation rests solely on the strength of association of the disablers and not on the number of accessible disablers.

Retrieving disablers from semantic memory will decrease the acceptance of MP and MT inferences. Therefore, we predict that acceptance ratings for MP and MT inferences will be higher for the strongly expanded conditionals than for the weakly ones. In the present experiment we did not manipulate the access to alternative conditions. Since, Cummins (1995) findings indicate that retrieving disablers has no effect on DA and AC it follows that no difference should be observed

on DA and AC acceptance between the strong and weak conditionals.

Experiment

Pretest

The material for the present experiment was selected from previous pilot work (see De Neys, Schaeken, & d'Ydewalle, 2000), where 20 participants wrote down as many disabling conditions as possible for a set of 20 causal conditionals (with 1.5 min generation time for each conditional).

For every conditional we established the relative frequency of appearance of the disablers that participants wrote down. We needed conditionals with a set of disablers in which there was one specific disabler that was very frequently generated. The expanded conditionals manipulation also forced us to take an additional criterion into account. We could not allow disablers that express a quantification of the original antecedent (e.g., 'brake not depressed hard enough'). Expanding the original with this kind of disablers would result in inconsistencies for some problems (e.g., DA, 'The brake was not depressed and the brake was depressed hard enough'). We selected 3 conditionals that met these criteria. From each set of disablers one infrequently generated disabler was selected. This weakly associated disabler had to meet the non-quantification criterion. Furthermore, if the strongly expanded conditional contained an explicit negation (e.g., 'If the apples are ripe and they are not picked'), we opted to express the selected weakly associated disabler in an explicit negated way too. The negation criterion should guarantee that the strongly and weakly expanded conditionals have comparable lexical complexity. Finally, the selected disablers had to sound as natural (according to our intuitions) as possible (e.g., 'not too little wind' was not accepted). Table 1 presents the material that was selected for the experiment.

We note that one might utter reservations about the use of frequency of generation as a measure of strength of association. Quinn and Markovits (1998) did not address this issue. However, our pilot study showed that frequency of generation was related to other possible strength of association measures such as plausibility and generation order: More frequently generated disablers were judged more plausible and tended to be generated prior to less frequently generated ones.

Method

Participants and Material

89 first-year university students participated in the experiment. Participants received a 4-page booklet. Page one included the instructions for the task. On the top of each of the next three pages appeared the selected conditionals. Each conditional was embedded in the four inference types (MP, DA, MT, AC). So, each of the three pages included one conditional with four inference problems. For each conditional there was a specific presentation order of the four inferences (AC, MT, DA, MP or MP, MT, DA, AC or MP, DA, MT, AC). The three pages were bound into booklets in randomized order. Below each inference problem appeared a seven point rating scale. This resulted in the following item format:

Rule: If water is heated to 100°C, then it boils

Fact: The water is heated to 100°C and the water is pure
Conclusion: The water boils

```
                                  I
-1--------2--------3---------4--------5--------6--------7-
very sure  somewhat sure  I somewhat sure  very sure
```

That I CANNOT draw this conclusion — That I CAN draw this conclusion

Figure 1. An example of the item format

Figure 1 presents an example of the MP problem. On the same page participants would also find the MT, DA and AC problem. The access to disablers manipulation

Table 1.
Relative frequency of generation of the most frequently mentioned disablers for the three selected conditionals. The disablers are given in order of frequency (%). Selected strongly and weakly associated disablers are highlighted.

If the apples are ripe, then they fall from the tree	If John grasps the glass with his bare hands, then his fingerprints are on it	If water is heated to 100° C, then it boils
Picked (65%)	*Hands not greasy (50%)*	*No pure water (75%)*
Too little wind (25%)	Grasped glass with palms only (35%)	*No normal pressure (30%)*
Not enough weight (20%)	Prints wiped off (30%)	Bad temperature measure (30%)
Not ripe enough (20 %)	*Glass was wet (25%)*	
Apples caught in branches (10 %)		

consisted in the presentation of two different minor premises (the information under the heading 'Fact'); the above example would belong to the strongly associated group were the original information was expanded with the negation of the strongest associated disabler. Similarly, in the weakly associated group, the negation of the selected weakly associated disabler was added to the 'Fact:'-information. In both expanded groups appeared the original conditionals on top of the item pages. Thus, participants were not presented explicit expanded conditionals but rather conditional inference problems with expanded minor premises. All the items in a single booklet belonged to the same group. Table 2 gives an overview of the different material in the two groups (for an MP problem)

Table 2.
Different contents in the experimental groups. Both groups only differ by the kind of information that is presented in the minor premise.

Expanded strongly associated:
(a) Water is heated to 100°C and the water is pure (b) The apples are ripe and they are not picked (c) John grasps the glass with his bare hands and his hands are greasy
Expanded weakly associated:
(a) Water is heated to 100°C and the pressure is normal (b) The apples are ripe and they are not caught in the branches (c) John grasps the glass with his bare hands and the glass is dry

Procedure
The booklets were randomly given out to students who agreed to participate in the experiment. No time limits were imposed. The instructions explained the specific item format of the task. Participants were told that the task was to decide whether or not they could accept the different conclusions. The instruction page showed an example problem (always standard MP) together with a copy of the rating scale. Care was taken to make sure that participants understood the precise nature of the rating scale. As in Cummins (1995), participants were NOT specifically instructed to accept the premises as always true. With Cummins we assume that this encourages people to reason as they would in everyday circumstances.

Results

Participants rated each of the four inference types three times. For every inference type the mean of these three ratings was calculated. This resulted in a 4 (inference type, within-subjects) x2 (group, between-subjects) design. All hypotheses were tested with planned comparison tests and rejection probability of .05.
Table 3 shows the overall mean acceptance ratings for the four inference types in the expanded weakly and strongly associated group. Planned contrasts indicated that the acceptance ratings in both groups differed significantly [$F(1, 87) = 4.55$, $MSe = 3.85$, $p < .04$]. As expected, both expanded groups did not differ in terms of the acceptance ratings for DA and AC inferences. For MT inferences we did obtain significantly higher ratings in the strongly associated group [$F(1, 87) = 4.99$, $MSe = 2.67$, $p < .03$]. Although, the effect on MP problems was in the expected direction (higher ratings in the strongly than in the weakly associated group), it did not reach significance.

Table 3.
Mean acceptance rating for the four inference types in the strongly associated and weakly associated groups.

Inference type	Group	
	Expanded weakly associated (n=45)	Expanded strongly associated (n=44)
MP	5.7	5.92
DA	4.78	5.11
MT	4.37*	5.14*
AC	4.98	5.44

* planned contrast $p<.05$

Discussion

The study showed that the strength of association of a disabling condition is affecting the conditional reasoning process. As predicted, peoples acceptance of MT inferences increased when there was no strongly associated disabler available, while the associative strength of the disablers had no effect on DA and AC inferences. This supports the hypothesis that in addition to the number of disabling conditions (Cummins, 1995), retrieving disablers from semantic memory is affected by their strength of association.

We suspect that the non-significance of the expected effect on MP may be due to a ceiling effect on the MP acceptance ratings. In the pretest, relatively few disablers were generated (less than the overall mean) for the three conditionals that were adopted for the experiment. Cummins (1995) already obtained high MP acceptance ratings for these conditionals. The 'expansion' manipulation in the present experiment

then further decreased the available number of disablers. This may have resulted in a ceiling effect. It could be the case that MP acceptance was already at the top in the weakly associated group. Mean acceptance for MP in the weakly associated group (Mean = 5.7, see Table 3) indeed tended to the 'sure that I can draw this conclusion' rating, located at the upper end of the scale. As in Cummins (1995), acceptance ratings on the (more difficult) MT inference were lower, what allowed the associative strength effect to show up.

It is interesting to note that for all of the four inference types acceptance ratings were lower when there was a strongly associated disabler available. Although the effect on AC and DA was not significant, one could suggest that the availability of a strongly associated disabler results in an overall decrease in certainty for every inference type (and not just for MT or MP). This might tie in with recent evidence (e.g. Manktelow & Fairley, 2000) showing that acceptance of DA can be affected by disabling conditions. This issue, together with the hypothesized ceiling effect on MP, will need to be addressed in further research.

In this study we adopted Quinn and Markovits' (1998) notion of a semantic search process and extended it to the disabling conditions. We should note that Quinn and Markovits (see also Markovits et al., 1998) incorporated the postulated semantic search process in the mental models theory (Johnson-Laird and Byrne, 1991). Here we refrained from making specific claims about the nature of the basic inferential principles (i.e., mental models or mental inference rules). The general semantic search process can be incorporated in other reasoning theories like mental logic (Braine & O'Brien, 1998; Rips, 1994) or the probabilistic approach (Oaksford & Chater, 1998). Comparing these different implementations is not within the scope of the present study or the Quinn and Markovits experiment.

We mentioned the relevance of the present study for the work of Elio (1997, 1998) and other researchers in the domain of belief revision and non-monotonic reasoning. Elio established that the number of stored disabling conditions affected peoples belief revisions and stated that conditional reasoning and belief revision are guided by the same memory search process. Our results show that successful retrieval is not only affected by the number of stored disabling conditions but also by their strength of association.

The present study can also be related to the work of Chan and Chua (1994). They examined the effect of 'relative salience' of disabling conditions. This factor can be interpreted as strength of association. Chan and Chua presented participants inference problems with two conditionals (e.g., 'If p then q, If r then q, p, thus q?'). The second conditional mentioned a possible disabling condition while the categorical premise was not expanded (see Byrne, 1989). Acceptance of MP and MT decreased with the strength of association of the mentioned disabler. However, a crucial difference with our study is that the present manipulation specifically affected the retrieval of disablers from semantic memory. In Chan and Chua's experiment, reasoning was affected by the strength of association of the mentioned disabler per se. The expansion of the categorical premise in the present experiment eliminated a strongly or weakly associated disabler and thereby affected the strength of association in the residual disabler set.

In sum, our study indicated that the conditional inferences people make are influenced by the strength of association of the disabling conditions. This complements Quinn and Markovits' (1998) contention that the strength of association of elements in semantic memory is an important factor in predicting conditional reasoning performance.

Acknowledgements

Preparation of the manuscript was supported by grants from the Fund for Scientific Research- Flanders (FWO) and the Belgian Program on Interuniversity Poles of Attraction, Convention number P4/19.

References

Braine, M. D. S., & O'Brien, D. P. (Eds.). (1998). *Mental logic*. Mahwah, NJ: Lawrence Erlbaum.

Byrne, R. M. J. (1989). Suppressing valid inferences with conditionals. *Cognition, 31*, 61-83.

Byrne, R. M. J., Espino, O., & Santamaria, C. (1999). Counterexamples and the suppression of inferences. *Journal of Memory and Language, 40*, 347-373.

Chan, D., & Chua, F. (1994). Suppression of valid inferences: Syntactic views, mental models, and relative salience. *Cognition, 53*, 217-238.

Cummins, D. D. (1995). Naive theories and causal deduction. *Memory and Cognition, 23*, 646-658.

Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory and Cognition, 19*, 274-282.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2000). *Causal conditional reasoning, semantic memory retrieval, and mental models: A test of the 'semantic memory framework'.* (Psychological report No.270). Leuven: University of Leuven. Laboratory of Experimental Psychology.

Elio, R. (1997). What to believe when inferences are contradicted: the impact of knowledge type and inference rule. *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society, 211-216*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Elio, R. (1998). How to disbelieve p -> q: Resolving contradictions. *Proceedings of the Twentieth Meeting of the Cognitive Science Society, 315-320*. Mahwah, NJ: Lawrence Erlbaum Associates.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M .J. (1993). *Human reasoning: The psychology of deduction*. Hove, UK: Lawrence Erlbaum.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Lawrence Erlbaum.

Manktelow, K. I., & Fairley, N. (2000). Superordinate principles in reasoning with causal and deontic conditionals. *Thinking and Reasoning, 6*, 41-65.

Markovits, H. (1984). Awareness of the 'possible' as a mediator of formal thinking in conditional reasoning problems. *British Journal of Psychology, 75*, 367-376.

Markovits, H., Fleury, M., Quinn, S., & Venet, M. (1998). The development of conditional reasoning and the structure of semantic memory. *Child Development, 69*, 742-755.

Newstead, S. E., Ellis, C. E., Evans, , J. St. B. T., & Dennis, I. (1997). Conditional reasoning with realistic material. *Thinking and Reasoning, 3*, 49-96.

Oaksford, M., & Chater, N. (1998). *Rationality in an uncertain world: Essays on the cognitive science of human reasoning*. Hove, UK: Psychology Press.

Quinn, S., & Markovits, H. (1998). Conditional reasoning, causality, and the structure of semantic memory : strength of association as a predictive factor for content effects. *Cognition, 68*, B93-B101.

Rips, L. J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.

Rumain, B., Connell, J., & Braine, M. D. S. (1983). Conversational comprehension processes are responsible for reasoning fallacies in children as well as adults. *Developmental Psychology, 19,* 471-481.

Thompson, V. A. (1994). Interpretational factors in conditional reasoning. *Memory and Cognition, 22*, 742-758.

Vadeboncoeur, I., & Markovits, H. (1999). The effect of instructions and information retrieval on accepting the premises in a conditional reasoning task. *Thinking and Reasoning, 5*, 97-113.

When knowledge is unconscious because of conscious knowledge and vice versa

Zoltan Dienes (dienes@biols.susx.ac.uk)
Experimental Psychology, Sussex University
Brighton BN1 9QG England

Josef Perner (Josef.Perner@sbg.ac.at)
Institut fuer Psychologie, Universitaet Salzburg
A-5020 Salzburg Austria

Abstract

This paper will offer a framework and a methodology for determining whether subjects have conscious or unconscious knowledge. The implicit-explicit distinction will be related to consciousness using the framework of Dienes & Perner (1999; 2001a,b,c) and the higher-order thought theory of Rosenthal (1986, 2000). Whether a mental state is conscious or not depends on whether certain inferences are unconscious or not, in a way we will specify; this is the interaction between implicit and explicit knowledge we will consider. The arguments will be illustrated with the artificial grammar learning paradigm from the implicit learning literature.

Introduction

In this paper we will argue that there is an intimate and generally unappreciated interaction between implicit and explicit knowledge that occurs all the time. Consideration of this interaction is important in determining whether a subject possesses conscious or unconscious states of knowledge. To make the argument, in the first section below we will overview the framework of Dienes and Perner (1999, 2001a,b,c) for understanding the implicit-explicit distinction in terms of the properties of representations. We will take an everyday use of the implicit-explicit distinction and apply it in a particular way to what it is to represent something. Given a representational theory of knowledge, this produces a hierarchy of ways in which knowledge can be implicit or explicit. We will then use the higher-order thought theory of Rosenthal (1986, 2000) to show full explicitness is almost the requirement for mental states of knowing to be conscious. There is one further stipulation above and beyond full-explicitness needed for consciousness and it is this that shows the importance of an interaction between implicit and explicit knowledge in producing conscious or unconscious states. We will discuss this relationship and illustrate how subjects' knowledge of an artificial grammar could be shown to be conscious or unconscious (in fact we will argue that the evidence shows that subjects can acquire fully unconscious knowledge).

Implicitly vs Explicitly Representing

Dienes and Perner's (1999, 2001a,b,c) framework could be structured as semi-independent modules: a notion of representation, a notion of the implicit-explicit distinction, the hierarchy of implicitness, and a theory of consciousness. To a degree, one can reject one of the modules and still accept the others to build an understanding of implicit knowledge and consciousness. We begin first with the notion of representation: In order to be clear how one might explicitly or implicitly represent something we need to be clear about what it is to represent something. In this we follow the functional theories of representation. For example, according to Millikan (1984, 1993), there must be a producer of the representation that has as its function that it brings about a mapping from the representation to a state of affairs. For example, in a bee dance, the bee can produce a dance such that the angle of the dance maps onto the location of the nectar. Further there will be consumers of the representation that perform various functions as they react to it. But they can only perform their functions under normal conditions if the representation does indeed map onto a certain state of affairs: This state of affairs is the content of the representation. This is what we mean by a representation. On this account, representations do not need to have further properties (e.g. compositional semantics) to be representations. The weights of a connectionist network are representations: They must map onto statistical regularities in the world in a certain way for the consumers of these weights to perform their functions, so the weights represent statistical regularities. What is it for a representation to represent something implicitly or explicitly, and what makes some representations conscious (those that have the contents that are the contents of our consciousness) while other representations are unconscious?

A bee dance represents the location of nectar. We say it represents location explicitly, because variations in the representational medium (angle of dance) map onto variations in location. However, it does not explicitly represent that it is about nectar: There is nothing in the medium that varies with whether it is nectar that it is about or something else. We say it represents the fact that it is about nectar only implicitly. Just so, in everyday terms when one answers the question "what is this?" to a succession of animals, and responds with the statement "cat" (or "dog" etc), the statement explicitly represents the property of being a cat, because variations in the representational medium (words) map precisely into variations in this content. The statement implies that it is this that is a cat, but it does not say so explicitly: There is not a part of the medium that varies directly with this rather than that being a cat.

Now consider what it is to have knowledge. In general knowledge consists of a proposition ("this word has the meaning butter") towards which you have an attitude of knowing. One can know without making all of these components of knowledge explicit. One can represent the proposition explicitly but not the fact that it is knowledge. It can in fact be knowledge (because it is taken as true and acted upon) without there being a representation that goes into one state for it being knowledge and another state if it is not knowledge. Minimally one could just make explicit the property without making explicit the individual that has this property (compare the bee dance). In subliminal perception we argue that it is indeed just a property of a presented word that is represented explicitly e.g. having the meaning "butter". There is no representation with the content "I see that the word in front of me has the meaning butter" so this cannot be the content of any experience of the subject; but the representation of merely "butter" can allow the subject to e.g. say "butter" as the first dairy product that comes to mind. At the next stage, the full proposition is made explicit ("The word in front of me has the meaning butter"). This stage involves the binding of features to individuals. At the next stage the factuality or otherwise of the proposition is made explicit ("it is a fact that the word in front of me has the meaning butter"). This is precisely the developmental milestone that occurs in a child's representational capacity at about 18 months (Perner, 1991), and is needed for appreciating hypotheticals, changing temporal states of affairs (and hence is necessary for explicit memory), and counter-factual reasoning. At the final stage the propositional attitude by which one holds the proposition is made explicit ("I see that the word in front of me has the meaning butter"). This is full self and attitude explicitness. We argue that this is necessary for knowledge to be conscious knowledge. This link from full explicitness to consciousness is made via the higher order thought theory of consciousness (Rosenthal, 1986, 2000; Carruthers, 1992, 2000).

The Higher-Order Thought Theory

Rosenthal (1986, 2000) develops an account of when a mental state is a conscious mental state. He argues that when one is in a conscious mental state one is conscious of that mental state. It would be inconceivable to claim that one is in a conscious state of, for example, seeing the word butter, while at the same time denying being conscious of seeing the word butter. So the question is, how does one become conscious of mental states? The relevant way, Rosenthal argues, is to think about them. We become conscious of our seeing the word butter when we think that we are seeing the word butter. That is, when we are consciously seeing the word butter, we have a thought like "I see that the word is butter". Because this thought (this mental state) is about another mental state (seeing), it is called a higher order thought. Note that this higher order thought is just our requirement for knowledge to be fully explicit: There is a natural relationship between explicitness and consciousness.

A Method For Determining Unconscious States

These considerations show that knowledge states being conscious or not is essentially a metacognitive issue (Dienes & Perner, 2001b). Roughly, simply knowing something and hence being able to respond discriminatively does not make the knowing a conscious mental state; for the latter, one must know that one knows. Metacognition has both a monitoring and a control aspect, and both of these aspects can be used to form methodologies for determining the conscious status of knowledge via an analysis of the relationship of different types of control and monitoring to the hierarchy of implicitness (Dienes & Perner, 2001b). Here we will focus exclusively on monitoring; the criterion for a state being conscious or unconscious is essentially that of the subjective threshold in the subliminal perception literature (Cheesman & Merikle, 1984).

Consider a subject in an artificial grammar learning experiment (Reber, 1967, 1989). The subject is exposed to strings of letters generated by a finite state grammar and asked to memorize them. After some minutes exposure, the subject is told actually there was a set of complex rules that determined the order of letters within the strings, and could they now classify a new set of strings as obeying the rules or not. Reber found that subjects could do so above chance but they found it difficult to say what the rules were. How could we

determine whether their knowledge is actually unconscious?

When subjects classify a test string they bring to bear their knowledge of the grammar to produce a new piece of knowledge: Whether this string is grammatical or not. We must distinguish these different knowledge contents: knowledge of the grammar, and knowledge of a particular string being grammatical (the grammaticality judgement).

When subjects make grammaticality judgements to the same strings several times they respond with different degrees of consistency to different strings (Reber, 1989; Dienes, Kurz, Bernhaupt, & Perner, 1997). For some strings the subject responds highly consistently, for others the subject may give a "grammatical" or "non-grammatical" response with 50% probability. Our interpretation of this fact is that subjects are in different knowledge states about the different test strings. Regardless of whether they have induced the same grammar as the experimenter or not (in fact, their grammar is correlated with the experimenter's grammar), the subjects themselves are treating themselves as being in different knowledge states about different strings. But have they conceptualized themselves as being in those different knowledge states? That is, have they formed attitude-explicit representations - higher-order thoughts - about those states? (Note that it was important to establish that there were different knowledge states before this question could be asked.)

When confidence ratings are taken after each classification decision, subjects can classify at above chance rates even when they claim they are literally guessing (for a review see Dienes & Berry, 1997). Further, under some conditions, there will be no within-subject relationship between confidence and accuracy: Subject do not know about the different knowledge states they are in fact in (Dienes & Berry, 1997; see Dienes & Perner, 2001c for this finding with a context-free grammar). Their knowledge is attitude-implicit, and hence unconscious.

The Crucial Interaction Between Conscious and Unconscious Knowledge

Consider now two objections one may have to assessing unconscious knowledge with confidence ratings. First, Allwood, Granhag, Johansson (in press) found the normal evidence for attitude-implicit knowledge in an artificial grammar learning experiment when the typical amount of learning and testing was used. In a second experiment that involved greater exposure to strings at learning and test subjects' confidence and accuracy was well-calibrated and so the knowledge seemed entirely atittude-explicit, despite the authors' feeling that the fundamental nature of subjects' knowledge had not changed. Allwood et al suggest that confidence ratings can come to be based on implicit knowledge as much as explicit knowledge, and this possibility undermines the usefulness of a dissociation between confidence and accuracy as a measure of implicit knowledge. Second, there is the standard objection to subjective measures of consciousness: Are they not dependent on the vicissitudes of subjects' idiosyncratic theories of consciousness, knowledge, etc (e.g. Shanks & St John, 1994)? This second objection is taken up in Dienes & Perner (2001a) and especially Twyman and Dienes (2001); we will develop a different, complementary response in this paper. To address these objections we need to return to a subtlety of the higher-order thought theory that we glossed over.

Rosenthal argues that to make a mental state conscious, the higher order thought must assert that one is in the state and it must not arise from any inferences of which we are conscious. That final stipulation is crucial. If I am driving along and swerve, and wonder why I swerved, I might think "I must have swerved because I saw that truck". That thought is a higher order thought about one being in a mental state of seeing. But it does not make the original seeing conscious; it does not make it conscious precisely because it arose from an inference of which one was conscious.

A mental state is conscious if we are non-inferentially conscious of it; the mental state would be unconscious if we were conscious of the mental state only by virtue of inferences of which we are conscious. This is just why blindsight patients' seeing is still unconscious even though they may consciously infer they must be seeing "because the experimenter tells me I am consistently correct." They have a higher-order thought that they are seeing, but our intuitions are that the seeing remains unconscious; and it remains unconscious precisely because the higher-order thought (the attitude explicit representation) arose from an inference of which the patient is conscious.

How do these considerations apply to determining whether a subject in an experiment has acquired conscious or unconscious knowledge? One has to be careful when the mental state is knowing, because knowing need not refer to an occurrent mental state at all. Knowing is often used in a dispositional sense: if you are asked if you know your times tables, that does not mean to say you are actively thinking about them now; the question is just whether you could do so accurately if asked. We will see how this can lead to a paradox.

Imagine a person is asked a general knowledge question and they believe they know the answer. The person is asked "Why do you conclude that you know

that?" When a person consciously knows something they might be able to provide conscious inferences by means of which they conclude that they know. They might justify their knowledge as knowledge because e.g. " I can describe the reasoning by which I drew the conclusion"; "I remember the event in which a trusted authority told me the knowledge"; or, more generally, "I can consciously link the knowing to some conscious perception."

These conscious inferences do not make the knowledge unconscious. (Further, they are not essential for the knowledge being conscious either: I might not know why I know something, I just insist that I know it.) This seems to go against the conclusion that mental states are only conscious if one knows about them by inferences of which one is unconscious.

The problem arises because knowing is not an occurrent mental state. An occurrent mental state associated with knowing is "thinking with conviction" or "thinking with a certain degree of conviction". Even if a person is aware of the inferences by which they know something, they just directly know that "I am thinking with conviction" if the thinking is a conscious state. In answer to the question "How do you know you are thinking with conviction?" one does not need to list the inferences that justify the knowledge as knowledge; they are not inferences leading to the conclusion that one is thinking with conviction. One just directly knows that one is thinking with conviction if the thinking-with-conviction is a conscious mental state. It is a conscious state because the inferences, if any, by which one ascertained that one was in the state were unconscious.

In answer to the question "Why do you conclude that you know that?" a person might provide conscious inferences by means of which they conclude they know something by observing their behaviour: "I respond consistently, quickly or effectively". For example, a person may select the correct capital of a country, let's presume, due to being in an unconscious state of thinking-with-conviction. This state makes the person respond consistently and quickly; the state of thinking-without-conviction (let's presume) makes the person respond inconsistently and slowly. The person does not know he is in a state of thinking with conviction at first; but he consciously infers from the speed with which the answer came to him that he must have been in an occurrent state of knowing. Because he is conscious of this inference, the state is unconscious. If the same inference had been drawn for the same reason but unconsciously, his thinking-with-conviction would be a conscious mental state. In this sense, explicit knowledge of ones mental state depends on that explicit knowledge being produced only implicitly; for example, only with inferences that are themselves implicitly represented. This is the crucial interaction between implicit and explicit knowledge we wish to dwell on.

Applying these notions to artificial grammar learning, consider a subject who sees a test string, applies knowledge-of-the-grammar, classifies the test string as grammatical or not (knowledge-of-the-test-string), and then gives a confidence rating.

Subjects' different degrees of consistency to different test strings show that subjects, when classifying different strings, are in states of thinking with different degrees of conviction. If subjects confidence ratings are unrelated to their consistency then their higher order thoughts (confidence ratings) are not sensitive to their actual mental states of thinking with more or less conviction ("knowledge states", for short). We have taken this to be evidence of the knowledge states being unconscious. In fact, however, in some cases they will be in a state of thinking-with-a-lot-of-conviction and give a high confidence rating; in these cases, they do have a higher-order thought (attitude-explicit knowledge) to the effect that they are in a state that they are in; so if the confidence rating came to them in a way that appeared unmediated, the state would be a conscious state. If confidence ratings appear unmediated to the subject, the lack of relationship between confidence and consistency implies some knowledge is unconscious, even though it allows some knowledge to be conscious. This is one refinement we must add to our previous interpretation of a lack of relationship between confidence and accuracy.

Now consider Allwood et al's results. Demonstrating a relationship between confidence and accuracy is not sufficient for demonstrating that the knowledge is conscious; one also needs to determine what the subject believes the confidence ratings are based on. Strictly speaking, if subjects base the ratings on inferences (e.g. perceived reaction times, perceived fluency) of which they are conscious, the knowledge states are still unconscious. Although we are in progress with an experiment that includes asking subjects to report on the bases of their confidence ratings, we regard this more as a means for us as psychologists to generate ideas, rather than as a test of the conscious status of their knowledge; the latter needs to be methodologically simpler. A lack of relationship between confidence and accuracy does imply that at least some of the knowledge states are unconscious, and so this remains a valuable criterion.

If subjects' ratings are based on explicit inferences, including the products of implicit knowledge (e.g. fluency), the knowledge states are unconscious; if confidence ratings come to be based on implicit

knowledge in a way that appears unmediated to the subject, and hence confidence is calibrated with accuracy, then the states of knowledge are conscious. The knowledge states referred to here are knowledge about the grammatical status of test strings. Even if these knowledge states were all conscious, it would leave open the possibility that knowledge of the grammar was unconscious.

Thus, Allwood et al's intuitions that subjects in their experiment two still had, in some sense, implicit knowledge could be due to (a) despite the subjects' well calibrated confidence ratings, this calibration was based on conscious inferences regarding the mental state the subjects must have been in (states of knowing the grammatical status of strings), and so those mental states were still in fact unconscious; or (b) the mental states of knowing the grammatical status of the strings were in fact conscious (the confidence ratings were not based on conscious inferences), but the mental states of knowing the grammatical rules were not conscious; subjects did not have non-inferential higher order thoughts about being in those latter states.

If subjects' become conscious of their mental states responsible for grammaticality judgements, methods of using confidence ratings for grammatcaility judgements can no longer be used to show knowledge of grammatcal rules is unconscious. Dienes and Perner (2001b) discuss how to measure implicit knowledge under these conditions.

Rosenthal (2000) discusses how higher order thoughts need not be produced by a 100% reliable means; however they are produced, so long as they appear unmediated, they produce conscious awareness of being in a certain mental state. The first order mental state about which one has a second order thought need not even exist for the subject to consciously experience being in a certain state. Subjects' higher-order thoughts partly constitute their theories about their mental states; such theories are not therefore a nuisance that get in our way as experimenters; they are part of the very thing to be investigated and explained.

Conclusion

We have argued for a second-order representational account of consciousness. Consciousness can never be produced just by, for example, a sustained pattern of activation in a connectionist network per se (e.g. O'Brien & Opie, 1999); the properties of the representation must in certain respects be like those of people to count as mental states and the system must be able to refer to those states in further representations (Perner & Dienes, 1999; see Dienes & Perner, 2001c for discussion). Such considerations lead directly to metacognitive measures of the consciousness of mental states. Conscious states cannot be measured just by discriminative responding, but only by evidence that the subject is conscious of the mental state.

To summarize the argument of this paper, we have taken two points made by Rosenthal regarding his higher order thought theory; namely that (a) higher order thoughts are occurrent states rather than dispositional states (like knowing might be); and (b) the higher order thoughts must not result from inferences of which the person is conscious. These two points turn out to have important implications for the measurement of implicit or explicit knowledge in (for example) the implicit learning literature. The contribution of this paper is to show in detail the relevance of Rosenthal's theory to psychologists interested in determining the conscious or unconscious status of mental states in (for example) implicit learning studies.

We are conscious of mental states when our explicit knowledge of them is based purely on implicit knowledge, when the knowledge of being in the state does not arise out of any inference of which we are conscious. Note there is a symmetry here with volition; we have voluntary control over an act only when the intention produces the act by mechanisms of which we are unconscious (Dienes & Perner, 2001b).

The importance of considering the conscious status of the inferences leading to a judgement has also been highlighted by Koriat (e.g. 1998). We hope we have elucidated further implications of the interaction between implicit and explicit inferences and implicit and explicit knowledge states.

Acknowledgments

Thanks to David Rosenthal for valuable discussion before this paper was written.

References

Allwood, C. M., Granhag, P. A., Johansson, H. (in press). Realism in confidence judgements of performance based on implicit learning. *European Journal of Cognitive Psychology*.

Carruthers, P. (1992). Consciousness and concepts. Proceedings of the Aristotelian Society, Supplementary Vol. LXVI, 42-59.

Carruthers, P. (2000). *Phenomenal consciousness naturally*. Cambridge: Cambridge University Press.

Cheesman J. & Merikle, P. M. (1984). Priming with and without awareness. *Perception & Psychophysics*, 36, 387-395.

Dienes, Z., & Berry, D. (1997). Implicit learning: below the subjective threshold. Psychonomic Bulletin and Review, 4, 3-23.

Dienes, Z., Kurz, A., Bernhaupt, R., & Perner, J. (1997). Application of implicit knowledge: deterministic or probabilistic? *Psychologica Belgica, 37*, 89-112.

Dienes, Z., & Perner, J. (1999) A theory of implicit and explicit knowledge. *Behavioural and Brain Sciences*, *22*, 735-755.

Dienes, Z., & Perner, J. (2001a) A theory of the implicit nature of implicit learning. In Cleeremans, A., & French, R. (Eds), *Implicit learning*. Psychology Press.

Dienes, Z., & Perner, J. (2001b). The metacognitive implications of the implicit-explicit distinction. In Chambres, P., Marescaux, P.-J., Izaute, M. (Eds), *Metacognition: Process, function, and use*. Kluwer.

Dienes, Z., & Perner, J. (2001c). Unifying consciousness with explicit knowledge. In Cleeremans, A. (Ed.) *The unity of consciousness: binding, integration, and dissociation.* Oxford University Press.

Koriat, A. (1998). Metamemory: the feeling of knowing and its vagaries. In M. Sabourin, F. Craik, & M. Roberts (Eds), *Advances in psychological science* (Vol. 2). Hove, UK: Psychology Press.

Millikan, R. G. (1984). Language, thought, and other biological categories. Cambridge, MA: MIT Press.

Millikan, R. G. (1993). *White queen psychology and other essays for Alice*. Cambridge, MA: Bradford Books/MIT-Press.

O'Brien, G., & Opie, J. (1999). A connectionist theory of phenomenal experience. *Behavioural and Brain Sciences, 22*, 127-196.

Perner, J. (1991). *Understanding the representational mind*.. Cambridge, MA: MIT Press. A Bradford Book.

Perner, J., and Dienes, Z. (1999) Higher order thinking. Behavioural and Brain Sciences, 22, 164-165.

Reber, A.S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behaviour, 6,* 855-863.

Reber, A.S. (1989). Implicit learning and tactic knowledge. *Journal of Experimental Psychology: General, 118*, 219-235.

Rosenthal, D.M. (1986). Two concepts of consciousness. *Philosophical Studies*, **49**, 329-359.

Rosenthal, D.M. (2000).Consciousness, Content, and Metacognitive Judgments, *Consciousness and Cognition*, 9, 203-214.

Shanks, D. R. & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioural and Brain Sciences*, 17, 367-448.

Twyman, M., & Dienes, Z. (2001). Metacognitive Measures of Implicit Knowledge. 2001 Convention of The Society for the Study of Artificial Intelligence and the Simulation of Behaviour (AISB), York March 21st-24th.

What Can Homophone Effects Tell Us About the Nature of Orthographic Representation in Visual Word Recognition?

Jodi D. Edwards (jdedward@ucalgary.ca)
Department of Linguistics, University of Calgary
2500 University Drive NW, Calgary, AB, Canada T2N 1N4

Penny M. Pexman (pexman@ucalgary.ca)
Department of Psychology, University of Calgary
2500 University Drive NW, Calgary, AB, Canada T2N 1N4

Abstract

In a lexical decision task (LDT), Pexman, Lupker, and Jared (2001) reported longer response times for homophones (e.g., MAID-MADE) than for non-homophones (e.g., MESS) and attributed these effects to orthographic competition created by feedback activation from phonology. The focus of the present research was the grain-size of the orthographic units activated by feedback from phonology. We created 9 categories of homophones based on the sublexical, orthographic overlap between members of homophone pairs. We also manipulated the type of foils presented in LDT (consonant strings, pseudowords, pseudohomophones) to create conditions involving less vs. more extensive processing. Homophone effect sizes varied by category; effects were largest when spellings of both onsets and bodies differed within the homophone pairs (e.g., KERNEL-COLONEL) and when members of the homophone pairs differed by vowel graphemes (e.g., BRAKE-BREAK). These results suggest that several specific grain-sizes of orthographic representation are activated by feedback phonology.

Introduction

In a number of recent articles in the word recognition literature, the notion of feedback activation has been invoked to explain particular findings (e.g., Hino & Lupker, 1996; Pecher, in press; Pexman & Lupker, 1999; Pexman, Lupker, & Jared, 2001; Stone, Vanhoy, & Van Orden, 1997; Taft & van Graan, 1998; Ziegler, Montant, & Jacobs, 1997). In a fully interactive model of word recognition (e.g., the PDP model of Plaut, McClelland, Seidenberg, & Patterson, 1996) activation between sets of units can be bi-directional. For instance, in a lexical decision task, when a target word is presented, there is initial activation of an orthographic representation for the target, and then very quickly there is also activation of semantic and phonological representations for that word. Those semantic and phonological representations then re-activate, via feedback connections, the orthographic representation. This bi-directional flow of activation can help the system settle on a representation for the target word. The purpose of the present research was to address an unresolved issue regarding feedback activation: What is the nature (grain-size) of the orthographic units that are activated by feedback from phonology?

Feedback activation is assumed to operate between all sets of units in the word recognition system. The focus of the present research, however, was feedback activation from phonology to orthography. Taft and van Graan (1998; see also Taft, 1991) argued for bi-directional activation between orthography and phonology by what they termed "orthography-phonology-orthography rebound". The model of word recognition they described was similar to models proposed by Grainger and Ferrand (1994), Plaut et al. (1996), and Van Orden and Goldinger (1994). A version of this model is illustrated in Figure 1. This is a connectionist model with sets of processing units representing orthographic, phonological, and semantic information. Importantly, the orthographic and phonological components of the model (but not the semantic component) are "composed of a hierarchy of units ranging from graphemes (e.g., C, A, and T) and phonemes (e.g., /k/, /æ/, and /t/) up to whole words. Activation passes up this hierarchy as well as between O and P units at the same level." (p. 206).

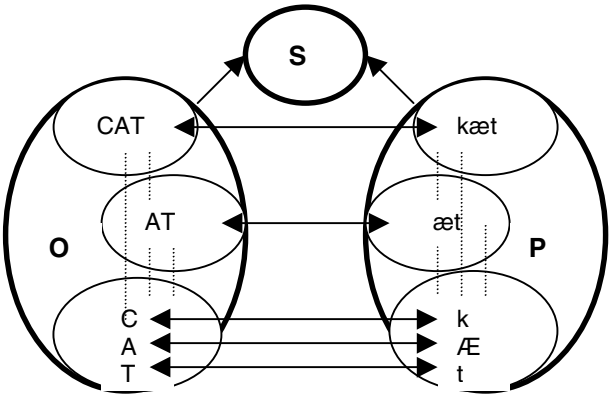


Figure 1: As depicted in Taft & van Graan (1998), a model of word recognition with sets of units representing orthography (O), phonology (P), and semantics (S).

Taft and van Graan (1998) argued that, when processing printed words, there is automatic activation of the phonological component of the model. Certainly, there has been controversy about the role that phonology plays in

visual word recognition (e.g., Davelaar, Coltheart, Besner & Jonasson, 1978; Jared & Seidenberg, 1991; Pugh, Rexer, & Katz, 1994, etc.). While Taft and van Graan found evidence that phonology did not mediate access to word meaning, they concluded that there was evidence for activation of phonology. That evidence came from studies involving homophone stimuli (e.g., MAID-MADE) (e.g., Jared & Seidenberg, 1991; Van Orden, 1987, etc.).

In several studies, researchers have investigated whether homophones create confusion when presented without context, in a lexical decision task (LDT) (e.g., M. Coltheart, Davelaar, Johnasson, & Besner, 1977; Rubenstein, Lewis, and Rubenstein, 1971). Recently, Pexman et al., (2001) reported longer decision latencies for homophones than for nonhomophonic control words in LDT, particularly for low frequency homophones with higher frequency homophone mates. Those homophone effects were larger with pseudohomophone foils than with pseudoword foils. Pexman et al. concluded that homophone effects in LDT were robust, and argued for automatic activation of phonology in visual word recognition.

Pexman et al. (2001) offered an account of homophone effects in LDT that was similar in many ways to the notion of orthography-phonology-orthography rebound (Taft & van Graan, 1998). Pexman et al.'s account was based on the concept of feedback phonology. The notion that feedback activation from phonology to orthography might influence the process of word recognition was first explored by Stone, Vanhoy, and Van Orden (1997).

Stone et al. (1997) argued that the process of word recognition is best explained by a model that includes both feedforward and feedback connections (resonance) between orthographic and phonological units. As support for this claim, they reported feedback consistency effects in LDT. Feedback inconsistent words are words for which the body can be spelled in more that one way (e.g., /-ADE/ in FADE can be spelled /-AID/ or /-AYED/ as in PAID or SWAYED), whereas feedback consistent words are words for which the body can only be spelled one way (e.g., /-IMP/ in LIMP). Stone et al. observed slower lexical decision latencies for feedback inconsistent words than for feedback consistent words. Accordingly, they suggested that, when a feedback inconsistent word is processed in an LDT, the phonological representation can activate, via feedback connections, orthographic representations for several word bodies. These orthographic representations will compete with each other, and this competition will slow the recognition process. In a replication study, Zeigler, Montant, and Jacobs (1997) also reported feedback consistency effects, thus supporting the notion that feedback activation can influence word recognition performance.

Although much current research supports the feedback account, it should also be noted that Peereman, Content, and Bonin (1998) reported a failure to replicate feedback consistency effects with French stimuli in LDT. Peereman et al. suggested that whereas homophone effects in various tasks "can be interpreted as showing that lexical phonological codes reverberate to orthographic word forms, they do not imply interactions between orthographic codes and phonological codes at the sublexical level" and argued for a "restricted interactivity account in which interactions are limited to lexical processing levels" (p. 170).

Pexman et al.'s (2001) feedback account of homophone effects in LDT also involved interactivity at the lexical level, although did not deny the possibility of sublexical interaction. Pexman et al. suggested that, when the phonological representation of a member of the homophone pair is activated, it feeds back to the orthographic representations for both spellings of the word. Thus, while some of the feedback is directed towards the representation for the correct orthographic unit, some of it is also captured by activity in the orthographic representation of the incorrect homophone mate, thus creating competition and resulting in longer decision latencies for homophones. This account involves the assumption that LDT responses are made primarily on the basis of activation in the orthographic units (see Pexman & Lupker, 1999, and Pexman et al., 2001, for more detailed explanations of this assumption).

The Present Research

If it is the case that phonology feeds back to, and creates competition in, the orthographic units, then it becomes important to characterize and attempt to understand the exact nature of this orthographic competition. In previous research there have been several different suggestions about the nature of orthographic representations. As mentioned above, one suggestion is that feedback activation from phonology influences activation of whole word orthographic representations (Peereman et al., 1998; Pexman et al., 2001). There have also been suggestions about the sublexical orthographic representations that might be activated by feedback from phonology. These sublexical representations have been described as grapheme based (Zeigler et al., 1997) or syllabically based, with feedback to onset and rhyme units (Stone et al., 1997), however, both Ziegler et al. and Stone et al. acknowledged that other levels of orthographic representation could be activated via feedback. There is, in fact, a suggestion that several different grain-sizes of units are activated in the process of word recognition, such that feedback activation would influence both lexical and sublexical levels of orthographic representation (Taft & van Graan, 1998). The purpose of the present research was to investigate the grain-size of orthographic units activated by feedback from phonology. To do this, we investigated whether homophone effects are modulated by the type of orthographic overlap that exists between homophone mates. In previous investigations of homophone effects in LDT (e.g., Pexman et al., 2001; Pexman & Lupker, 1999; Rubenstein et al., 1971) homophone pairs have only been categorized by frequency. Yet homophone pairs vary widely in orthographic overlap; some homophone pairs differ only by a single internal grapheme (e.g., BERTH-BIRTH), while others differ by onset and also by word body (e.g., ATE-EIGHT). Our question for the present research was whether these sublexical differences in orthographic overlap lead to

differences in the size of observed homophone effects. To this end, we created nine separate categories of homophones. Our aim was to divide homophones according to types of sublexical, orthographic overlap within homophone pairs (see Table 1), but our divisions between categories were also unavoidably influenced by the type of homophones that tend to occur in English. We restricted our analysis to low frequency homophones that have higher frequency mates since these were the homophones that produced the largest effects in Pexman et al. (2001).

In the following experiments we presented low frequency homophones from each homophone pair in the above homophone categories, along with sets of low frequency non-homophonic control words matched to the low frequency homophones. Our tasks were 3 LDTs, across which we manipulated the type of foils presented, to create task conditions that required less vs. more extensive processing. In Experiment 1A foils were consonant strings (e.g., PRNVR), in Experiment 1B foils were pseudowords (e.g., PRANE), and in Experiment 1C foils were pseudohomophones (e.g., BRANE). Pexman et al. (2001), and Pexman and Lupker (1999) have reported that when foils are more word like (e.g., pseudohomophones), homophone effects are larger. The explanation is that pseudohomophone foils create a difficult LDT, in which participants tend to process all of the stimuli more extensively. With more extensive processing, there is more opportunity for feedback activation to influence activation at the orthographic level and, hence, more competition and larger homophone effects. By using progressively more difficult LDTs, we hoped to capture homophone effects at several different "moments" of processing, allowing for more thorough contrasts between the homophone categories.

Method

Participants

The participants in these experiments were undergraduate students at the University of Calgary. There were 35 participants in Experiment 1A, 37 participants in Experiment 1B, and 41 participants in Experiment 1C.

Stimuli

Words The words used in this experiment were 95 homophones (mean frequency = 16.92 per million; Kucera & Francis, 1967) and 95 control words (mean frequency = 15.43) matched for frequency, onset, length and neighbourhood size (Coltheart, Davelaar, Jonasson, & Besner, 1977).

Foils Foil stimuli were required in all three parts of the experiment. There were 95 foils of each of the three types: consonant strings (Experiment 1A), pseudowords (Experiment 1B), and pseudohomophones (Experiment 1C).

Procedure

On each trial, a letter string was presented in the center of a 17-inch Sony Trinitron monitor controlled by a Macintosh G3 and presented using PsyScope (Cohen, MacWhinney, Flatt & Provost, 1993). Lexical decision responses were made by pressing either the left button (labelled NO) or the right button (labelled YES) on a PsyScope response box.

Experiment 1A – Results and Discussion

For this and each of the following experiments, mean decision latencies, mean error percentages, and homophone effect sizes for each category are presented in Table 1. In all analyses, data were analyzed with subjects ($F1$ or $t1$) and, separately, items ($F2$ or $t2$) treated as random factors.

To test the view that whole-word units are the important orthographic units for feedback activation, we conducted a 9 (homophone category type) X 2 (homophony) ANOVA to see if the effects of homophony varied by category. The overall homophone effect was significant in the latency analysis ($F1(1, 34) = 6.73$, $p < .05$, $MSE = 3514.11$; $F2(1, 86) = 4.26$, $p < .05$, $MSE = 916.53$), and in the error analysis ($F1(1, 34) = 18.87$, $p < .001$, $MSE = 128.50$; $F2(1, 86) = 18.86$, $p < .001$, $MSE = 36.66$). Thus, we confirmed the existence of homophone effects in LDT, replicating the results of Pexman et al. (2001), but here with a larger set of items and with consonant string foils. There was a main effect of category in the latency analysis ($F1(8, 27) = 3.81$, $p < .01$, $MSE = 3422.01$; $F2(8, 86) = 2.37$, $p < .05$, $MSE = 1514.47$) and in the error analysis ($F1(8, 27) = 5.08$, $p < .01$, $MSE = 54.31$; $F2(8, 86) = 1.90$, $p = .07$, $MSE = 46.84$). The interaction of category and homophony was not significant in the latency analysis ($F1(8, 27) = 1.71$, $p = .13$, $MSE = 4446.21$; $F2(8, 86) = 1.47$, $p = .18$, $MSE = 916.53$) but was significant in the error analysis by subjects ($F1(8, 27) = 5.19$, $p < .01$, $MSE = 38.05$; $F2(8, 86) = 1.71$, $p = .11$, $MSE = 36.66$). These effects indicate that the size of the homophone effects differed somewhat across the nine categories of homophones. Since this LDT involved consonant string foils, decisions could be made on the basis of relatively shallow processing.

As illustrated in Table 1, none of the homophone effects were significant in both latency and error analyses. Significant latency effects were observed for the Body Only and Onset and Body categories, and significant error effects were observed for the Single Vowel Only, Silent E or Word Internal Dipthong, /-s/ Morpheme, and Silent Onset and Body categories. In the case of the Single Vowel Only and Silent Onset and Body categories, error rates were relatively high for the homophones (15.0 % and 10.0 %, respectively). These error rates are surprisingly high for a LDT involving consonant string foils, and suggest that some of our participants may not have known some of these homophones (e.g., BERTH, WHOLLY, etc.). In the latency analyses for these categories, which include only correct responses, there were no differences between latencies for homophones and latencies for control words. Thus, the error effects in these categories may not really be indicative of orthographic competition. Hence, in the following experiments we draw conclusions only about homophone effects that are significant in both latency and error analyses.

Table 1: Homophone Effect Sizes

Homophone category	Example	Experiment 1A (consonant string foils) RT	Error	RT effect	Error effect	Experiment 1B (pseudoword foils) RT	Error	RT effect	Error effect	Experiment 1C (pseudohomophone foils) RT	Error	RT effect	Error effect
Single vowel only	berth	541	15.0			642	28.1			742	33.9		
Control	blink	548	3.4	-7	+11.6* **	597	6.9	+45*	+21.2* **	692	10.5	+50*	+23.4* **
EA or EE grapheme only	deer	511	4.9			562	9.2			675	14.6		
Control	deed	512	3.6	-1	+1.3	551	6.7	+11	+2.5	621	4.7	+54*	+9.9*
Silent E	brake	531	4.9			589	10.2			698	14.9		
Control	bleed	505	2.3	+26	+2.6	539	5.2	+50* **	+5.0*	601	2.0	+97* **	+12.9*
Silent E or word internal dipthong	maid	522	7.0			581	11.6			682	15.4		
Control	mess	518	1.9	+4	+5.1*	522	1.5	+59* **	+10.1*	600	2.2	+82* **	+13.2* **
/-ed/ morpheme	guessed	537	4.3			576	4.8			693	5.6		
Control	glimpse	549	4.5	-12	-0.2	600	9.6	-24* **	-4.8*	682	8.4	+11	-2.8
/-s/ morpheme	presents	537	7.9			594	16.0			724	16.7		
Control	pleasant	537	2.9	0	+5.0*	578	6.5	+16	+9.5*	681	7.7	+43*	+9.0*
Body only	suite	534	3.7			577	7.4			684	8.2		
Control	shirt	518	1.8	+16**	+1.9	545	1.8	+32* **	+5.6* **	623	3.3	+61* **	+4.9*
Onset and body	kernel	558	7.1			619	15.9			726	19.1		
Control	kennel	516	3.9	+42* **	+3.2	560	5.4	+59*	+10.5*	633	4.0	+93* **	+15.1* **
Silent onset and body	wholly	540	10.0			599	17.7			709	24.1		
Control	wildly	532	5.0	+8	+5.0*	584	8.9	+15	+8.8*	669	8.3	+40	+15.8*
Foils		499	2.0			640	6.1			732	6.4		

*$\underline{p} < .05$ by subjects, **$\underline{p} < .05$ by items

Experiment 1B – Results and Discussion

In the 9 (homophone category type) X 2 (homophony) ANOVA, the overall homophone effect was significant in the latency analysis ($\underline{F}1(1, 36) = 43.01$, $\underline{p} < .001$, $\underline{MSE} = 3118.18$; $\underline{F}2(1, 86) = 21.57$, $\underline{p} < .001$, $\underline{MSE} = 2966.35$), and in the error analysis ($\underline{F}1(1, 36) = 75.17$, $\underline{p} < .001$, $\underline{MSE} = 139.16$; $\underline{F}2(1, 86) = 18.09$, $\underline{p} < .001$, $\underline{MSE} = 150.30$). The main effect of category was significant in the latency analysis ($\underline{F}1(8, 29) = 5.55$, $\underline{p} < .001$, $\underline{MSE} = 3147.14$; $\underline{F}2(8, 86) = 2.52$, $\underline{p} < .05$, $\underline{MSE} = 6521.49$) and in the error analysis ($\underline{F}1(8, 29) = 8.39$, $\underline{p} < .001$, $\underline{MSE} = 77.04$; $\underline{F}2(8, 86) = 2.19$, $\underline{p} < .05$, $\underline{MSE} = 178.82$). Importantly, the interaction of category and homophony was significant in the latency analysis ($\underline{F}1(8, 29) = 5.74$, $\underline{p} < .001$, $\underline{MSE} = 3007.84$; $\underline{F}2(8, 86) = 2.34$, $\underline{p} < .05$, $\underline{MSE} = 2966.35$) and was significant by subjects in the error analysis ($\underline{F}1(8, 29) = 11.69$, $\underline{p} < .001$, $\underline{MSE} = 74.91$; $\underline{F}2(8, 86) = 1.80$, $\underline{p} = .09$, $\underline{MSE} = 150.30$). These effects indicate that the size of the homophone effects differed across the nine categories of homophones. This result confirms that homophone effects vary by category and reveals that the source of homophone effects is not only competition from whole-word orthographic units. Thus, it is not the case that homophone effects arise whenever one phonological representation maps onto two orthographic representations. The magnitude of homophone effects seems to depend to some extent on competition between orthographic units that represent the sublexical structure of the homophones.

As illustrated in Table 1, homophone effects were observed in this experiment for 5 of the 9 types of homophones. The largest homophone effects seemed to arise in the categories of homophones that differ from their high frequency mates in onset structure (as long as the onset is articulated, since there was no effect in latencies for the Silent Onset and Body category), body structure, or by a single vowel grapheme (although not for the EA or EE Grapheme category).

Notably, the /-ed/ Morpheme homophones actually produced an effect in the reverse direction, while the /-s/ Morpheme homophones produced a null effect in the latency analysis. A tentative conclusion is that homophone effects do not arise for homophones that differ in morphological structure from their homophone mates (e.g., GUESSED-GUEST). Before interpreting this result any

further, we examined effect sizes for all categories again in Experiment 1C.

Experiment 1C – Results and Discussion

As in Experiment 1B, the overall homophone effect was again significant in the latency analysis ($F1(1, 40) = 77.58$, $p < .001$, $MSE = 9481.57$; $F2(1, 86) = 35.62$, $p < .001$, $MSE = 7658.91$), and in the error analysis ($F1(1, 40) = 170.92$, $p < .001$, $MSE = 138.34$; $F2(1, 86) = 27.68$, $p < .001$, $MSE = 212.49$). There was a main effect of Category in the latency analysis ($F1(8, 33) = 7.93$, $p < .001$, $MSE = 7307.81$; $F2(8, 86) = 2.24$, $p < .05$, $MSE = 13926.27$) and in the error analysis ($F1(8, 33) = 24.87$, $p < .001$, $MSE = 87.94$; $F2(8, 86) = 2.59$, $p < .05$, $MSE = 235.16$). There was also an interaction of Category and Homophony that was significant by subjects in the latency analysis ($F1(8, 33) = 2.74$, $p < .01$, $MSE = 7082.59$; $F2 < 1$) and in the error analysis ($F1(8, 33) = 13.20$, $p < .001$, $MSE = 87.14$; $F2\ (8, 86) = 1.54$, $p = .16$, $MSE = 212.49$). These effects indicate that the size of the homophone effects differed across the nine categories of homophones.

As in Experiment 1B, the greatest homophone effects in Experiment 1C were observed in the categories where homophone pairs differed by a vowel grapheme or by onset-body units. These effects affirm the notion that graphemes and onset-body units are important sources of competition for homophones. The implication is that these units receive feedback activation from phonology.

The /-ed/ Morpheme category demonstrated facilitation for homophones in Experiment 1B, yet, in Experiment 1C with pseudohomophone foils, we found that this facilitation disappeared and a null homophone effect was observed instead. This finding, along with the relatively small homophone effects observed for the /-s/ Morpheme category across foil conditions, suggests that homophone effects are not generally observed for pairs of homophones that have different morphological structure. There are two possible interpretations of these null homophone effects. One interpretation is that the orthographic representations for the two members of the homophone pairs are so similar that no competition arises. The second interpretation is that the orthographic representations for the two members of the homophone pairs are so different that no confusion or competition arises. We would tend to support the latter interpretation. If one ignored morphological structure, the homophones in the /-ed/ Morpheme category could be classified as Body Only homophones. Yet the Body Only homophones produced quite robust homophone effects compared to those produced by the /-ed/ Morpheme homophones. Therefore, the morphological structure of the /-ed/ homophones is an important factor in explaining the null (and sometimes facilitory) effects for that category. Homophones like GUESSED, that have different morphological characteristics than their homophone mates (GUEST), are apparently not confused with their homophone mates at the orthographic level. The extra morpheme /-ed/ seems to create an orthographic representation that is easily distinguished from the orthographic representation for the homophone mate.

As in Experiment 1B, the homophone effect for the Silent Onset and Body category in Experiment 1C was not significant in the latency analysis (although it was significant by subjects in the error analysis). Again, there are two possible interpretations for a null (or relatively small) homophone effect. One possible interpretation is that the representations for the words in these homophone pairs are so similar that minimal competition arises. That is, 'silent' letters may not have much bearing on the nature of orthographic representations for words like WHOLLY or KNOT. Hence KNOT may be encoded very much like NOT, with little competition arising. The second interpretation is that the representations for the words in these pairs are so different that minimal competition arises. That is, because the onsets and many of the bodies are orthographically different within the homophone pairs, KNOT may be easily distinguished from NOT, resulting in minimal competition. We tend to favour the first interpretation. The reason for this is that the homophone effect for the Onset and Body category is much larger than the homophone effect for the Silent Onset and Body category. The fact that the effect size is markedly smaller for the Silent Onset and Body category suggests that the silent onsets are not competing in the same way that the articulated onsets are, causing smaller (non-significant) homophone effects.

General Discussion

The purpose of the present research was to conduct a precise examination of the orthographic factors that modulate homophone effects, in order to determine the grain size of units activated by feedback activation from phonology.

The homophone effects observed in the experiments reported here provide support for the notion that phonology is activated in the process of visual word recognition and feeds back to units in orthography. We also observed differences in the extent to which different types of homophones produced homophone effects. Analysis of effect sizes for our homophone categories revealed that homophone effect sizes varied by sublexical orthographic overlap of homophone mates. Homophone effects were greatest when the members of homophone pairs differ by a single vowel grapheme, or by the word body, or by the word body and articulated onset, within one morpheme boundary. In terms of identifying precisely what the levels of sublexical representation are, the trends in our data suggest that the levels likely correspond to graphemes, and onsets and bodies. We acknowledge, however, that the ambiguities inherent in the orthography of English homophones (e.g., some of our homophones differed slightly on orthographic properties other than those defined by the category labels, many of our categories were "grapheme" categories since these are the most common type of English homophone pairs) prevent us from making stronger conclusions. Nonetheless, the cross-category differences in our homophone effects make it apparent that the feedback

process does not reflect a mapping of phonology onto only whole word constituents at the orthographic level. This is not to say that lexical units are not also involved in the feedback process. According to the model depicted in Figure 1, activation at sublexical levels within the orthographic units feeds up to the lexical level. Presumably, activation at the lexical level must reach a certain point before a response is made. For homophones, responses seem to be delayed by competition at sublexical levels within the orthographic units. These delays are most obvious when an LDT is difficult (e.g., with pseudohomophone foils), because a higher threshold of activation is set and hence competition must be more fully resolved before a response is made.

These data provide support for a fully interactive model of word recognition, in which sublexical information is part of the orthographic and phonological components (e.g., Plaut et al., 1996; Taft & van Graan, 1998; Van Orden & Goldinger, 1994). The homophone effects reported in this paper suggest that there is bi-directional activation between the orthographic and phonological components of such a model, and that this activation is captured in several different grain-sizes of representation.

Acknowledgements

This research was supported by a summer studentship from the Alberta Heritage Foundation for Medical Research (AHFMR) to the first author and a research grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) to the second author. We thank Lorraine Reggin and Gregory Holyk for assistance testing participants.

References

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, & Computers, 25,* 257-271.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI*. Hillsdale, NJ: Erlbaum.

Davelaar, E., Coltheart, M., Besner, D., & Jonasson, J. T. (1978). Phonological recoding and lexical access. *Memory and Cognition, 6,* 391-402.

Grainger, J., & Ferrand, L. (1994). Phonology and orthography in visual word recognition: Effects of masked homophone primes. *Journal of Memory and Language, 33,* 218-233.

Hino, Y., & Lupker, S. J. (1996). Effects of polysemy in lexical decision and naming: An alternative to lexical access accounts. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 1331-1356.

Jared, D., & Seidenberg, M. S. (1991). Does word identification proceed from spelling to sound to meaning? *Journal of Experimental Psychology: General, 120,* 358-394.

Kucera, H., & Francis, W. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Pecher, D. (in press). Perception is a two-way junction: Feedback semantics in word recognition. *Psychonomic Bulletin and Review.*

Peereman, R., Content, A., & Bonin, P. (1998). Is perception a two-way street? The case of feedback consistency in visual word recognition. *Journal of Memory and Language, 39,* 151-174.

Pexman, P. M., & Lupker, S. J. (1999). Ambiguity and visual word recognition: Can feedback explain both homophone and polysemy effects? *Canadian Journal of Experimental Psychology, 53,* 323-334.

Pexman, P. M., Lupker, S. J., & Jared, D. (2001). Homophone effects in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 139-156.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*, 56-115.

Pugh, K. R., Rexer, K., & Katz, L. (1994). Evidence of flexible coding in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 807-825.

Rubenstein, H., Lewis, S. S., & Rubenstein, M.A. (1971). Evidence for phonemic recoding in visual word recognition. *Journal of Verbal Learning and Verbal Behavior, 9,* 487-494.

Stone, G. O., Vanhoy, M., & Van Orden, G. C. (1997). Perception is a two-way street: Feedforward and feedback phonology in visual word recognition. *Journal of Memory and Language, 36,* 337-359.

Taft, M. (1991). *Reading and the mental lexicon*. Hove, England UK: Lawrence Erlbaum Associates, Inc.

Taft, M., & van Graan, F. (1998). Lack of phonological mediation in a semantic categorization task. *Journal of Memory and Language, 38,* 203-224.

Van Orden, G. C. (1987). A ROWS is a ROSE: Spelling, sound, and reading. *Memory & Cognition, 15,* 181-198.

Van Orden, G. C., & Goldinger, S. D. (1994). Interdependence of form and function in cognitive systems explains perception of printed words. *Journal of Experimental Psychology: Human Perception & Performance, 20,* 1269-1291.

Ziegler, J.C., Montant, M., & Jacobs, A.M. (1997). The feedback consistency effect in lexical decision and naming. *Journal of Memory and Language, 37,* 533-554.

Memory Representations of Source Information

Reza Farivar (reza@ego.psych.mcgill.ca)
Department of Psychology, McGill University
Montreal, QC, H3A 1B1, Canada

Noah Silverberg (noahs@uvic.ca)
Department of Psychology, University of Victoria
Victoria, BC, V8P 5C2, Canada

Helena Kadlec (hkadlec@uvic.ca)
Department of Psychology, University of Victoria
Victoria, BC, V8P 5C2, Canada

Abstract

Various characteristics can be encoded to define the source of particular information. How these characteristics interact to describe and define a source has so far been ignored. Our work focuses on the representation of source information in memory using the General Recognition Theory. The results are discussed in relation to current modelling efforts.

Introduction

How is a source defined? How do we represent context information about an event? How do we use this information in source monitoring? Source monitoring refers to the proper attribution of an item to its source (Lindsay and Johnson, 1991; Johnson, Hashtroudi, and Lindsay, 1993). Sources may be defined by a variety of characteristics, including but not limited to size, font, location, cognitive processes at encoding, affect, and so on. For written material, font, location, size, and possibly the syntax and structure used in a part of text may identify a part of text as a source. Memory for source is vulnerable, and source misattributions (SM) are common. Such errors may arise due to (1) failure of cognitive processes underlying the judgment, (2) adoption of a lax criterion which does not involve deliberate and conscious consideration of each judgment, and (3) similarity of the sources to such an extent that they are indistinguishable from one another. In other words, source monitoring involves normal memory processes and decisional processes, and a separation of the two can yield a better understanding of how sources are actually defined and how decisional processes can fail and cause SM errors.

Recently, a number of multinomial models of source monitoring have been proposed (Bayen, Murnane, and Erdfelder, 1996; Batchelder and Reifer, 1990; see Batchelder and Reifer, 1999 for a review). These statistical models assume a processing tree with a single root. Each branch of the tree represents one possible path in the stage of the process, and each fork in the branch represents a possible division of probabilities for a specific outcome. The branches end in response categories, from which all the fork parameters are estimated. This type of modelling allows an analysis of every possible contribution to each factor and thus enables one to model a cognitive process into discrete stages, collect categorical data, and then estimate the contributions made at each stage.

The advantages of such a model are easily visible. Interpretations are made effortless by the simple tree structure. Assumptions of independence between various fork parameters allow one to divide and separate various components of a process and to demonstrate independence between these components. However, such representation of source information is not modelled by this technique, and as such, these modelling efforts are limited in capturing the global features of source monitoring.

There are at least two major limitations to multinomial modelling. The first involves the degrees of freedom in the estimation procedure. Given that there are often more parameters than categories, some parameters must be assumed equal. This may sometimes be beneficial, since a model is reduced to its minimal, simplest components. Unfortunately, even with this parameter reduction, one may still have more parameters than categories, and the estimation procedure will thus result in a number of possible models rather than a single possible solution. From this pool of resulting models, using goodness of fit tests, one must determine which model actually represents the expected design or the data.

A consequence of this is that as findings in the field increase, such a model is more difficult to grow because of the increase in the free parameters. A multinomial model will therefore be limited to more general processes and will not be able to model details of a process, such as how source information is represented in memory. A lack of understanding about the details of a process can, in turn, result in a misunderstanding of the process as a whole as well

as the relation of that process in the grand scheme of cognition.

The second limitation, already hinted at, is the requirement of assuming statistical independence between various fork parameters. What this means is that a process stage, modelled on a fork further down a tree, is thought to be independent of a process stage modelled further up the tree; no two processes can have dependence in such a model. As we will see below, this assumption presents a crucial limitation of multinomial models in capturing the representation of source information in memory.

How to test statistical independence

In a multidimensional setting of a source characteristics where various factors influence source memory and source judgments, how is one to assess independence while accounting for the influence of all memory and decision factors? Historically, for unidimensional stimuli, signal detection theory (SDT; Banks, 1970; see Swets, 1996) has been the method of choice for separating decisional and perceptual factors in perception and recognition. However, SDT does not allow for a test of independence and is also limited to the analysis of only unidimensional stimuli. Ideally, we would want a statistical technique similar to SDT that could also account for interactions of various dimensions on a recognition task while providing a test of independence and separation of decisional and perceptual factors. General Recognition Theory (GRT; Ashby and Townsend, 1986; Ashby, 1988), and the analytical method permitted by it (Multidimensional Signal Detection Analysis, or MSDA, Kadlec & Townsend, 1992a; Kadlec & Townsend, 1992b), meet our requirements.

GRT and MSDA

The GRT was developed by Ashby and Townsend (1986) in response to various issues that had been raised in perception research. These issues concerned the notion of independence and separability of the perception of stimulus dimensions as well as decisional factors. The GRT is a formal method of assessing independence and separability in terms of both stimulus dimensions and decisional processes. MSDA was then developed by Kadlec and Townsend (1992a) in order to facilitate the implementation of GRT in perception studies. This analytic method maps the traditional SDT parameters of sensitivity (d') and response bias onto a multidimensional scheme and permits us to analyse both interaction and independence between various stimulus dimensions.

Multidimensional signal detection analysis MSDA was originally developed to analyse the effects of manipulation on the perception of a stimulus, when the manipulations are varied on a number of dimensions (Kadlec, 1995). An example study would be one that looks at the dependence of the perception of eyebrow curvature on the perception of lip curvature. First, we would need a feature-complete factorial design, and we would create this by manipulating each of the two dimensions on two levels. We would thus manipulate each of the two dimensions on two levels. Thus the eyebrows will be varied on two levels (low and high curvature) and the same manipulation would be used for lip curvature. Table 1 demonstrates this matrix for all variations of stimulus A.

Table 1: Example of a feature-complete factorial design.

	Eyebrow Curve	
Lip Curve	Low $_{(a)}$	High $_{(b)}$
Low $_{(i)}$	A_{ai}	A_{bi}
High $_{(j)}$	A_{aj}	A_{bj}

The participant will see stimuli, which vary on two levels of two dimensions. It should be noted that the stimuli at each level vary only slightly from the stimuli at the other level. This is mainly due to the fact that the final analysis will be based on the amount of errors committed during the task, and simplifying the task and making the difference between the stimuli obvious may result in a near absence of errors.

In a typical experiment utilizing this design, participants are given a practice session in which they view the set of stimuli, one at a time, varied on the given dimensions, and make judgments about their location at the different levels of the dimensions. In other words, a participant who views stimulus Abj must try to categorize this stimulus on both dimensions. The correct response for this stimulus is high eyebrow curvature and high lip curvature. There are three types of errors that could be made in the categorization of each stimulus. These errors can be tabulated, resulting in proportions of each type of error, which then represents the volume under the distribution in one of 4 possible response regions. These response regions can be represented by normal distributions in multiple dimensions.

How could we, from these data and types of distributions, answer questions regarding independence of dimensions and decisional boundaries? In order to answer such questions, a slightly different view of the graphs must be utilized. Consider a plane passing horizontally through all the normal distributions in this multidimensional space at a given density level. Examining such a plane from above would yield a topography of the distributions, as can be seen in Figure 1.

The shape of these distributions corresponds to three types of independence (Ashby & Townsend, 1986). The first type is perceptual independence (PI), which is a statistical form of independence. PI is stated when for a stimulus Abj:

$$fbj(x,y) = gbj(x)*gbj(y)$$

In this equation g(x) refers to marginal densities, which are obtained by integrating (measuring the area under the curve) the two-dimensional density distribution across one dimension. Marginal densities can be thought of as the picture of a density distribution as would be taken from having a camera parallel to a dimensional axis.

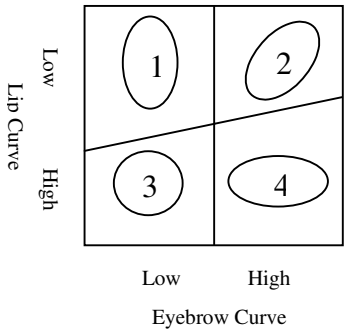


Figure 1: Topography of the distributions

PI is a strictly statistical form of independence and can be likened to coin toss probabilities, where the probability of obtaining two heads (assuming a fair coin; p(1H) = 0.5) is equal to the product of the probability of obtaining one head by the same probability:

$$p(2H) = p(1H) * p(1H) = 0.25$$

Thus PI asserts that the perception of one dimension within a stimulus is not dependent on the perception of the other dimension. From the topographical diagram, PI is represented by circular distributions and distributions which are elliptical but parallel to the axis of one dimension. From the diagram it may be observed that within stimuli 1, 3 and 4 the two dimensions are perceptually independent and the two-dimensional density distributions are equal to the product of the marginal densities of the stimulus on the individual dimensions.

Another form of independence, called perceptual separability (PS), is taken to exist when within one level of one dimension, the levels of the other dimension do not affect perception. In this case,

$$gi1(x) = gi2(x) \text{ and } gj1(x) = gj2(x)$$

where 1 and 2 refer to stimuli 1 and 2, i and j represent the two levels of eyebrow curvature, and g(x) refers to their marginal density distributions. If the two marginal distributions at the levels of one dimension (say high eyebrow curvature) are equal, then levels of the other dimension (lip curvature) do not influence the perception of the eyebrow curvature.

A third form of independence is decisional separability (DS). Recall that within the single dimension signal detection framework, a decisional boundary was set between the two distributions. Here too, within the multidimensional framework, some decisional boundary must be set. This decisional boundary is set by the participant and defines the area within which a stimulus will be identified by its specific characteristics. In our example, a decisional boundary must be set in order to differentiate between faces that vary differentially on eyebrow curvature and lip curvature. In other words, the decisional boundaries divide the multidimensional space into regions that define specific stimuli. Within this context, DS is observed when the decision about one dimension is not influenced by the decision made on the other dimensions. In our topography, DS holds when the decisional boundaries are parallel to the dimensional axis.

Methodology

From the above description of MSDA and its required experimental paradigm, it is more apparent how we would go about testing the independence of various source characteristics in a final source judgment. What we need is to create a feature-complete factorial design of at least two stimulus characteristics that can define a source, and then conduct our analysis on these dimensions.

Our study uses the characteristics of written text to assess this independence. The dimensions of written text tested are limited to size (large vs. small) and location (top vs. bottom). Within a multinomial framework, these two dimensions would have to be assumed independent of one another. In other words, independent of decisional biases, memory for an item being on top should have no effect on the memory of it being large. Such a model would represent a very basic and simple framework in which all sources are believed to be equal; if source characteristics do not interact and are independent, then any combination of the characteristics is remembered equally well.

Below we will present data to show that in this case, the assumption of independence is invalid, thus lying outside the structure of a multinomial model of source memory. Because the question of how various characteristics interact to define a source is central to the concept of source memory, we propose that multinomial models are limited in that they cannot capture this fundamental aspect of source memory.

Experiment 1

Participants

Eighteen undergraduate students enrolled in an introductory psychology course at the University of Victoria participated in the study for bonus marks. Responses from 3 participants were excluded because of incompleteness due to shortness of time. Responses from an additional participant were excluded due to apparatus failure. Overall, data from 14 participants were analysed. All data were kept confidential and no one was penalized for non-participation.

Material

A word-list was composed using 256 five-letter words randomly selected from the Francis-Kucera Frequency Norms. Of the 256 words, 160 were used at study and the remaining words were used as controls (novel words). Study items were factorially manipulated on size (large or small), font (Times or San Sarif), and location (top or bottom). An IBM-compatible computer with an Intel 486 processor was used for the experiment. All presentations were made on a 17" computer monitor and responses made on the computer keyboard.

Procedure

Participants were informed that they would view a list of words on the computer screen, varying in location, size, and font. They were instructed to try and remember all aspects of the words, as their memory for the words as well as these attributes would later be assessed. Following the instructions, the study list of 40 words was presented, each for 3 seconds, with 1second inter-stimulus-interval.

Immediately following the study phase, the test phase was conducted. Participants viewed the previously studied words in addition to 16 new words not previously seen. The words appeared in a small, neutral font (courier) at the centre of the screen. The first task was to make a remember/know/new judgement.

Following the instructions on remember/know/new judgements, participants were instructed on the identification task. Questions assessing the recognition for various item characteristics followed the remember/know/new judgements. The three questions asking for the recognition of levels of the three dimensions were randomised in presentation, so that for one item, font was asked first, followed by location and size, but for another, size was asked first followed by location and font, etc.; the sequence was randomised.

For both the remember/know/new judgements and the item judgements, participants made their response by selecting the first letter of the response on a computer keyboard (‘ r’ for ‘remember’, ‘t’ for times font when making font decision and top when making location decisions, and so on). The task was not timed, and participants were encouraged to consider each judgement carefully. Following the first study and test phase, three more study and test phases continued with the same instructions.

Results

Responses were collapsed across all subjects. To minimize variability due to practice (at the beginning of the session) as well as variability due to fatigue (at the end of the session), only responses from trials 2 and 3 were analysed; responses from these two trials were then combined, after no significant difference between them was observed on a chi-squared test. Due to space limitations, the results on remembering and knowing will not be discussed here, but we will only comment that near identical interactions were observed for remembered and known items; for the analysis, results associated with “remember” and “know” judgments (i.e. studied items judged as “old”) were combined.

The results are reported following Kadlec (1995). All tests of DS, PS, and PI are Z-tests with at least $p < 0.1$, unless otherwise noted. DS and PS held for all dimensions, whereas PI failed for all items.

To simplify the interpretation, the results were collapsed on to two dimensions, size and location. The resulting distribution topographies are represented in figure 2.

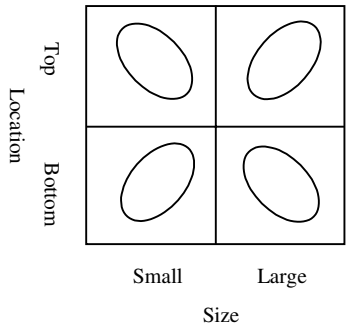


Figure 2: Schematic representation of results from experiment 1

The straight vertical and horizontal lines separating the distributions represent the decisional boundaries; they are parallel to the dimensional axese, because DS held for both dimensions. Furthermore, the distances between the distributions do not differ between the levels of each dimension, thus PS is also valid. However, PI has been violated in all instances, and in a consistent pattern: there is a strong positive tendency to remember large items as being on top and small items on the bottom, while there is a strong negative tendency to remember small items on top and large items on the bottom. Clearly, source definitions formed from these characteristics will result in a memory bias independent of decisional processes such as response biases. In other words, recolllected large words have a “topness” associated with them, and are dissociated from a “bottomness”, while the opposite is true for small words.

In order to further validate our results, we conducted a similar study, but we eliminated the presentation of a font dimension all together, as well as the remember/know judgments.

Experiment 2

Participants

Twenty-one undergraduate students enrolled in an introductory psychology course at the University of Victoria participated in the study for bonus marks.

Materials

One hundred sixty words with a frequency rating of 1 were obtained from the Francis-Kucera Word Pool and randomly divided into four lists of forty words. List presentation was fully counterbalanced. All the words were presented in random order in a study phase and were factorially manipulated on size (large and small) and location (top or bottom).

Procedure

Participants were informed that they would view a list of words on a computer screen, and that these words would vary on two dimensions, location on screen and font size. They were then instructed to do their best to 'remember' the words and how they were presented, as they would later be tested on these attributes. Firstly, in phase 1, a list of 40 words was presented for approximately 2 seconds each with an inter-stimuli interval of 1 second. Following this was a test phase in which subjects were asked 3 questions about 80 words (40 studied and 40 novel words): "Is this word old (o) or new (n)?", "Did it appear on the top (t) or bottom (b)?", and "Was it presented in a large (l) or small (s) size font?". The order of the questions was randomised. Subjects responded using a keyboard with the abbreviated letters corresponding to the responses to which they stand for (e.g. (o)=old). This task was not timed. Phase 2 followed and was a mere repetition of phase 1 but with different word lists. The whole session lasted 40 to 50 minutes.

Results

All tests of DS, PS, and PI are two-tailed Z-tests with at least $p<0.05$, unless otherwise noted. Analysis of the compiled old-correct matrix (where subjects responded "old" and were correct in doing so, whether in phase 1 or phase 2, or whether the new/old question preceded or followed the attribution questions) revealed an interaction of source characteristics, identical to that of experiment 1, represented in figure 2. Whereas DS and PS held for the two dimensions, PI failed in every case, such that there was both a positive dependency between "bottomness" and "smallness" and a positive dependency between "topness" and "largeness". Alternatively, it is also true that as words appear closer to the top of the screen, they are more likely to be remembered as having appeared in a larger size font. In analysis of unremembered items (where the item was presented, but judged as "new"), a bias was observed for words presented on top of the screen and in large size. This suggests that people have poor memory for such items; this and other result will be discussed below.

Discussion

How are source characteristics defined then? Using MSDA, we were able to separate decisional factors from memory factors. Our results are thus not based on a any decisional biases of the participant. In other words, it is not the case that participants recall an aspect of an item (e.g. having been presented in large size) and then infer from this information that the word must have been presented on top. If it were the case that participants recalled information on one dimension and inferred information on the other dimension, different criterion measures (C's) would have been observed at each level of any one dimension. However, measures of C were identical at every level of all dimensions, suggesting that decisional factors did not play a role at generating the observed patterns.

Perceptual separability also held for both dimensions, which suggest that memory for large items is no better than memory for smaller items, and that memory for items on top is no better than memory for items on the bottom. If this were not the case, d' measures for items presented on top or bottom would have differed when varied across the two sizes, and/or d' measures for items presented in large or small size fonts would have differed across the two different locations.

As both DS and PS held for both dimensions, we can infer that (1) decisional factors did not play a role in producing the observed pattern of results, and (2) that global features of the stimuli did not play such a role either. It can therefore be concluded that the interaction of the various source characteristics is a pure memory process.

This has implications for how source information is stored in memory. Our results suggest an information-compression taking place during encoding and/or storage of the source characteristics. In effect, error for the overall storage of such source information (size and location) is minimized by reducing error on the most frequently occurring instances of such a source. Source definitions composed of large items presented on the top of the screen are particularly common in all printed media such as newspapers, magazines, web pages, etc. Meanwhile, it occurs rarely that small print appears on the top part of a display, and the same is true for large items presented on the bottom of the screen. It may be the case that, in order to minimize errors on these frequently occurring source types, we are introducing biases into the process.

We believe that the compression (or consolidation) account is the most appropriate for our data; i.e. the way new information is stored is affected and directed by prior knowledge. This is analogous to perceptual illusions: perceptual illusions may exist as a consequence of a biased set of inputs which then results in a biased perception of illusory, but-otherwise-neutral items. It is likely that the same processes that cause perceptual illusions are responsible for the effects obtained here. In essence, our results may be demonstrative of a form of memory illusion, very distinct from prior studies of memory illusions with relation to eyewitness memory. In the case of eyewitness memory, some researchers (i.e. Loftus, Miller & Burns, 1978) suggest that presentation of secondary, post-event information impairs memory for an event by altering its contents. We suggest that *prior knowledge* impairs what new information

can be learned in the first place, by limiting how the new information is represented.

What is the relation between our results and multinomial models of source monitoring? We suggest that multinomial models will be unable to incorporate our data in a meaningful way, because of inherent limitations in the estimation procedure for such models. Here, we are referring to the fact that multinomial models would have to assume independence between various source characteristics and, as our results show, such assumptions would be invalid. Therefore, such modelling techniques will not be able, in the long run, to present an accurate model of source memory.

Our results do not falsify any present uses of multinomial models. Multinomial modelling is insensitive to biases for sources, thus such modelling could capture the more general nature of our results. However, such a technique inherently does not allow for any type of representation of source memory and we believe that such a representation is crucial for an accurate model of source memory.

Recall from our results that studied items that were judged as "new" were more often items that had been presented on top of the screen and in large font. Although statistical analysis of this trend was not conducted for this paper, this may be a real effect worth considering, as it suggest that certain sources (or certain combinations of source characteristics) are easier to remember, such that this source combination may even affect item detection. Again, as statistical analysis of this was not conducted, we are only speculating on this point, but it is consistent with our findings, as they suggest a form of memory consolidation taking place.

Future work in the area can thus focus on assessing how sources are defined, as the tools now exist to answer such specific questions. Modelling efforts should also attempt to incorporate this representational information, as this will, without doubt, contribute to a more complete understanding of human memory.

Acknowledgements

This research was supported in part by a grant to the third author from the National Science and Engineering Research Council of Canada (NSERC). We thank Aidan Quigley for his help in collecting data.

References

Ashby, F. G. (1988). Estimating the parameters of multidimensional signal detection theory from simultaneous ratings on separate stimulus dimensions. *Perception & Psychophysics*, 44, 195-204.

Ashby, F. G. & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, 93, 154-179.

Bayen, U. J., Murnane, K., & Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22, 197-215.

Batchelder W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modelling. *Psychonomic Bulletin & Review*, 6, 57-86.

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review*, 97, 548-564.

Johnson, M. K., Hashtroudi, S., and Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, 114, 3-28.

Kadlec, H. (1995). Multidimensional signal detection analyses (MSDA) for testing separability and independence: A Pascal program. *Behaviour Research Methods, Instruments, & Computers*, 27, 442-458.

Kadlec, H., & Townsend, J. T. (1992b). Signal detection analyses of dimensional interactions. In F. G. Ashby (Ed.*), Multidimensional models of perception and cognition*. Hillsdale, NJ: Erlbaum.

Kadlec, H., & Townsend, J. T. (1992a). Implications of marginal and conditional detection parameters for the separability and independence of perceptual dimensions. *Journal of Mathematical Psychology*, 36, 325-374

Lindsay, D. S., & Johnson, M. K. (1991). Recognition memory and source monitoring. *Bulletin of the Psychonomic Society*, 29, 203-205.

Loftus, E. F., Miller, D. G. & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory,* 4, 19-31.

Swets, J. A. (1996). *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Testing Hypotheses about Mechanical Devices

Aidan Feeney
Department of Psychology
University of Durham
Science Laboratories
South Road
Durham DH1 3LE
United Kingdom
aidan.feeney@durham.ac.uk

Simon J. Handley
Centre for Thinking and Language
Department of Psychology
University of Plymouth
Drake's Circus
Plymouth PL4 8AA
United Kingdom
shandley@plymouth.ac.uk

Abstract

We use the literature on mechanical reasoning to derive predictions about how people will test a mechanical rule. In the presence of a single rule we predict significantly more selections of tests in which the hypothesized cause is manipulated than in the presence of two rules: the original and one casting doubt on the sufficiency of the hypothesized cause for the effect. We describe an experiment using Wason's selection task that confirms our predictions and go on to discuss the implications of our results for recent work on causal cognition.

Introduction

Much of our everyday reasoning concerns the operation of mechanical devices. Many of our interactions with such devices, from figuring out how to program the VCR to mastering the newest piece of software, require us to draw inferences, make predictions and test hypotheses. Some of this reasoning is very important and, accordingly, mistakes can be costly. For example, the Chernobyl explosion (see Medvedev, 1991) has been used by Johnson-Laird (1993) to illustrate how biased interpretation of evidence relevant to an hypothesis about a device can lead to disaster. Fortunately, not all faulty mechanical reasoning leads to such calamitous consequences and sometimes we even have the luxury of experimenting with a device in order to see how it works. For example, the paradigmatic case in the literature on discovery learning (Khlar & Dunbar, 1988) is how people acquire the ability to operate a device by self-guided trial and error learning.

In this paper we will be concerned with the principles that determine the tests that people choose to carry out in order to examine an hypothesis about a mechanical device. Consider, for example, a conditional hypothesis concerning the cooling system in a factory:

(1) If the valve is open then water flows through the pipe.

Our concern is how people will go about testing such a rule rather than how they should test the rule. There are four cases which seem intuitively relevant here: the valve being open; the valve being closed; water flowing through the pipe; and water not flowing through the pipe. According to a normative theory based on the hypothetico-deductive method (Popper, 1959), in order to test the rule participants should chose to discover information about what happens both when the valve is open and when there is no water running through the pipe. Only these tests can provide evidence that falsifies the rule. Much research suggests that people do not consider these to be the best tests with which to examine a hypothesis and we do not think the case of mechanical reasoning is likely to be any different in this respect.

We suspect, however, that causal hypotheses concerning a mechanical device - such as the one above - will elicit a very specific pattern of testing behaviour amongst participants. The detail of our suspicions and a more comprehensive rationale will be presented in the next section.

Mechanical Reasoning

There is a variety of work on mechanical reasoning in the literature all of which suggests that people use representations that are in some way analog to the device being reasoned about. For example, Hegarty (1992) has proposed an account of how people think about devices claiming that mechanical reasoning processes are isomorphic to the structure of the device that is the subject of those processes. Hegarty's theory was intended to explain results obtained using a mechanical reasoning paradigm where participants are shown static images of pulley systems and are asked to predict the effect of some manipulation of the device. For example, participants might be asked how one or other of the pulleys in the image will move when a rope is pulled. Hegarty's data, both from eye-movements and gross reaction times, suggest that in making inferences about a mechanical device, people animate contiguous elements of the system piecemeal, by inferring a causal chain of events from the input of the system to its output.

Other work supports this conclusion. Schwartz & Black (1996), in showing that people can induce rules to describe a system based on mental depictions of that system, provide evidence suggesting that the initial depictions are constructed in a causal order. That is, in the absence of a rule, people appear to mentally

simulate the effects of a manipulation to the system in the direction of cause and effect.

People's piecemeal animation (Hegarty, 1992) and depiction (Schwartz & Black, 1996) of mechanical devices, suggest that their representations of mechanical devices respect the causal structure of the device. Furthermore, when people perform thought experiments involving those representations, they perform them so as to observe the effects on the system of the manipulation of a causal agent.

Testing Mechanical Hypotheses

We assume that in testing a causal hypothesis concerning a mechanical device people will form a representation of the hypothesized rule that respects its causal direction. So, given hypothesis 1:

If the valve is open then water will flow through the pipe

we expect participants to encode in their representation the information that, under this hypothesis, the valve is of causal significance with respect to water flowing through the pipe.

Given previous work on how people animate and perform mental experiments on models of devices, we expect participants, when imagining the possible consequences of performing a test on the system described in the hypothesis, to represent the consequences of the antecedent condition being in a certain state. That is, we would expect participants to be more interested in tests that respect the causal structure of the hypothesis than in tests that require backwards reasoning from changes in the effect to changes in its putative cause. Specifically, we expect participants to be more interested in cases where the cause is present or absent than in cases where the effect is present or absent.

The prediction that participants will be interested in cases where the effect is absent is a risky one, as participants are not normally interested in the false antecedent case when testing a conditional rule. Indeed, Oaksford and Chater's (1994) probabilistic account of how people test conditional rules claims that the false antecedent case is never informative. The situation for causal conditional rules is very different, however, where interest in the false antecedent case might be interpreted as being due to the use of a counterfactual strategy in testing the causal status of the antecedent. Mackie (1974) claims that we infer causality not only from repeated observations of contiguous events but also from a consideration of what might be observed in the absence of the putative cause. If the effect is also absent under these circumstances then we infer a causal relationship between the two. Harris, German & Mills (1996) have demonstrated such a strategy in the causal reasoning of children aged between 3 and 5 years.

There are, however, conditions under which we would not expect participants to be primarily interested in tests that respect the causal structure of the device. For example, if the hypothesis is presented at the same time as a second rule:

(1) If the valve is open then water will flow through the pipe
(2) *If the pipe is free from blockages then water will flow through the pipe*

where this second rule specifies an additional antecedent for the consequent, then we would expect participants to select fewer tests where the hypothesized cause is manipulated. This is because the additional antecedent introduces a potential disabling condition (Cummins et al 1991) for the hypothesized cause. This would mean that a failure to find the effect in the presence or absence of the cause might be attributable either to the hypothesized cause being insufficient to produce the effect or to the absence of the enabling condition. In the example above, the valve being open and the pipe being free from blockages might be conjointly necessary for water to flow through the pipe. If this is the case then examining the results of tests involving manipulation of the valve is unlikely to be revealing of the truth or falsity of the rule in the absence of information about the presence or absence of the enabling condition.

A Mechanical Selection Task

To test our intuitions about how causal rules about a mechanical device are represented and hence tested, we constructed a mechanical version of Wason's selection task (Wason, 1968). In our version of the task participants received a scenario (see below) which supplied a context for a conditional rule describing a causal relationship between the state of a component of the device and some output. Underneath were printed four cards representing the true antecedent, the false antecedent, the true consequent and the false consequent states of affairs. To test our hypotheses concerning the conditions under which participants would be primarily interested in tests of the hypothesis that manipulated the putative cause, we constructed a second version of the task. This second version was achieved by adding a second rule to the problem specifying an additional antecedent for the same outcome (see 1 and 2 above).

Our manipulation of number of rules is directly analogous to the presentation of additional antecedents in the conditional arguments task (Byrne, 1989). Participants who receive conditional reasoning problems that specify an additional antecedent are significantly less likely to draw the valid Modus Ponens and Modus Tollens conclusions than are participants who do not receive information about an additional

antecedent. This may be interpreted as the result of the additional antecedent causing participants to doubt the sufficiency of the first antecedent for the rule.

Byrne's work was an extension of experiments reported by Rumain, Connell & Braine (1983) who showed that presenting participants with a second conditional rule that specified an alternative antecedent for the consequent suppressed the rate at which the invalid Denial of the Antecedent and Affirmation of the Consequent inferences were drawn. Recently, Feeney and Handley (2000; Handley, Feeney & Harper, 2000) have described the results of selection task experiments where participants received a second rule specifying an alternative antecedent for the consequent in the rule to be tested. Across a series of six experiments large and reliable rates of suppression of Q card selections were found. A meta-analysis of five of the six experiments showed that the rate of not-P card selection was also significantly lower in the presence of an alternative antecedent.

The Q and not-P cards on the selection task are logically equivalent to the DA and AC inferences. Suppressing the rate at which they are selected is analogous to suppressing the rate at which the invalid inferences are made. One prediction about our mechanical selection task, therefore, is that the presence of a second rule specifying an additional antecedent should produce suppression on those cards which are logically equivalent to the MP and MT inferences i.e. the P and not-Q cards. However, we predict that this will not be the case.

Instead we expect rates of antecedent card selection (both true and false antecedent cards) to decrease in the presence of an additional antecedent. This prediction is based on our assumption that in reasoning about an hypothesis concerning a mechanical device people will incorporate into their representation the hypothesized causal status of the antecedent with respect to the consequent. Based on previous work (Hegarty, 1992; Schwartz & Black, 1996) we hypothesize that they will prefer to examine the results for the effect of manipulating the antecedent rather than vice versa. Such a testing strategy should be significantly reduced in the presence of a second rule containing an additional (and perhaps conjointly necessary) antecedent.

Method

Participants: 90 female and 21 male students at the University of Durham participated in this experiment. Participants' mean age was 20.5 years and age ranged from 17 to 42 years.

Design, Materials and Procedure: There were two groups of participants in this experiment. The first received a one-rule selection task containing a scenario and just one rule. The other group of participants received a two-rule selection task comprising of the same scenario and rule to be tested as well as a second rule specifying an additional antecedent for the consequent in the first rule. For all participants the scenario and the rule to be tested were as follows:

> *A friend of yours, who works in a factory, takes you on a tour of her place of work. She points to a large pipe and says that the cooling system in the factory obeys the following rule:*
>
> **If the valve is open then water will flow through the pipe**
>
> *You are interested in checking whether the cooling system does follow the rule your friend has told you about. Below are four cards which refer to tests that have been carried out on the cooling system. On one side of each of these cards is recorded whether the valve was open when the test was carried out whilst on the other side is recorded whether the water was flowing at the time of the test. Please indicate by circling the appropriate card or cards, which one(s) you need to turn over to decide whether the rule is true or false.*
> *Remember the rule you are testing is:*
>
> **If the valve is open then water will flow through the pipe**

The second rule received by half of the participants was as follows:

> **If the pipe is free from blockages then water will flow through the pipe**

Participants in this latter group were reminded that their task was to test the first rule. Finally, all participants saw four cards labelled 'Valve open', 'Valve closed', 'Water flowing' and 'Water not flowing'. They were asked to indicate those card(s) which were necessary in order to decide whether the rule was true or false.

Results

We performed three analyses on our results. The first examined the effects of experimental condition on individual card selection frequencies whilst the second was of the effect of number of rules on the rate at which antecedent and consequent cards were selected. In addition, we analyzed the frequency of various card combination selections.

Individual Card Selection Frequencies: Our first analysis was of the effects of our number of rules manipulation on the rate at which individual cards were selected in the experiment. These rates are presented in Table 1 below.

Chi-square analyses showed no significant effects of number of rules on the rate at which any of the cards were selected (P card: $\chi 2(1) = 1.73$, $p > .18$; not-P card: $\chi 2(1) = 1.40$, $p > .23$; Q card: $\chi 2(1) = 1.96$, $p > .16$; not-Q card: $\chi 2(1) = .15$, $p > .70$). As the presence of a

second rule has previously been found to significantly affect the total number of cards selected by participants (Feeney & Handley, 2000) we tested for an effect of our number of

Table 1: Percentage of participants selecting each card as a function of condition.

	P	Not-P	Q	Not-Q
One-Rule	84	40	37	21
Two-Rules	74	30	50	24

rules manipulation on the total selected in this experiment. This effect was not significant ($t(109) = .365$, $p > .71$). The mean total for the one rule condition was 1.82 cards (S.D. = .60) and 1.78 cards (S.D. = .74) for the two-rules condition.

Cause vs. Effect Selections: To test our predictions about cause and effect card selections we analysed the rate of cause and effect selections. For the purposes of this analysis we computed the number of cause and effect selections made by each participant, where a cause selection was defined as the selection of either of the antecedent cards and an effect selection as choosing either of the consequent cards. The mean numbers of each type of selection are shown in Figure 1 below.

We performed a 2x2 mixed design Anova on the cause and effect data. The between participants variable in this analysis was number of rules whilst the within participants factor was the number of selections concerning the hypothesized cause vs. the number of selections concerning the hypothesized effect specified in the rule. This analysis produced a non-significant main effect of number of rules ($F(1, 109) = .133$, MSE = .227), a highly significant main effect of cause vs. effect ($F(1, 109) = 33.40$, MSE = .385, $p < .001$) as well as a significant interaction between the within and between participant variables ($F(1, 109) = 4.94$, MSE = .385, $p < .028$). Tests for simple effects showed that

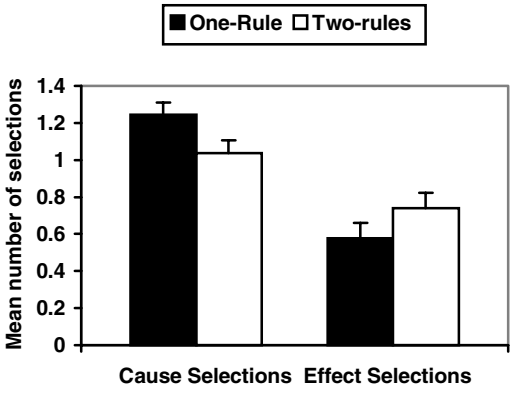


Figure 1: Mean number of cause and effect selections by condition.

there were significantly more antecedent (or cause) selections in the one-rule condition than in the two rule condition ($F(1, 109) = 4.964$, MSE = .243, $p < .028$). The difference due to number of rules on the rate of consequent (or effect) selections was not significant ($F(1, 109) = 1.965$, MSE = .369, $p > .16$).

Finally, in order to analyze the effect of our number of rules manipulation on relative rates of cause and effect card selections, we computed an index for each participant of their total of cause selections minus their total of effect selections. The effect of the number of rules manipulation on this index was significant ($t(109) = 2.22$, $p < .029$). In the one rule condition the mean score on this index was .667 (S.D. = .932) whereas in the two rule condition the mean score was .296 (S.D. = .816).

Card Combination Frequencies: Our final analysis was of the card combination frequency data from the experiment. As may be seen from Table 2, the rate of logically correct responding was not affected by our experimental manipulation with 4 out of 57 participants choosing the logically correct combination in the one-rule condition versus 5 out of 54 in the two-rule condition. The most striking effect of our number of rules manipulation on the combinations of cards that participants choose to select concerned the combined selection of the P and not-P cards only. In the one-rule condition 13 participants (23%) chose this combination whereas in the two-rule condition it was chosen by only 4 participants (7%). This difference is statistically significant ($\chi 2$ (1) = 5.07, $p < .02$) and is in the direction suggested by our hypothesis concerning how people represent and test causal rules concerning mechanical devices.

Table 2: Card combination frequencies as a function of condition.

	Condition	
Combinations	One Rule N = 57	Two Rules N = 54
p	13	13
not-p	2	1
q	0	5
not-q	0	0
p, not-q	4	5
p, not-p	13	4
p, q	14	14
not-p, q	2	2
not-p, not-q	4	5
q, not-q	1	1
p, not-p, not-q	1	1
p, q, not-q	2	0
all four	1	3

Discussion

The results of this experiment support our predictions about how people test causal rules concerning mechanical devices. Across both conditions we found a significantly greater rate of antecedent than consequent selection. This is not surprising and is probably true of most selection task experiments (although for an important exception, see below). Of much more immediate interest is the finding that the difference between the rate of antecedent and consequent card selections was significantly greater in the one-rule condition than in the two-rule condition. In addition, the rate at which people selected antecedent cards was significantly greater in the one-rule condition than in the two-rule condition although the (non-significant) rate of suppression due to the presence of a second rule was the same for each antecedent card. Finally, participants in the one-rule condition were significantly more likely to select the combination of cause present and cause absent cards than they were in the two- rule condition.

We argue that this pattern of results suggests that people's representation of the rule contains information about the putative causal status of the antecedent and when considering the possible consequences of the various tests of the rule people primarily consider tests where the hypothesized cause is manipulated. When compared to the results of selection task experiments where participants are asked to test a standard indicative rule, this experiment may be seen to have produced a very high rate (40%) of not-P card selection in the one rule condition. For example, in the meta-analysis reported by Handley and Feeney (2000) of single-rule conditions from five experiments on the standard indicative selection task, 24% of the 272 participants whose data were included were found to have selected the not-P card. We argue that the elevated rate of not-P selection found in this experiment is due to the use of a counterfactual strategy to test the causal claim made in the experimental rule. In addition to testing for the presence of the effect in the presence of the putative cause, participants were interested in looking to see whether the effect was present or absent when the hypothesized cause was absent.

Strikingly, when an additional antecedent calls the sufficiency of the putative cause into question, people are significantly less likely to select antecedent cards. A reduction in P card selections is expected given that the second rule may lead people to question the sufficiency of the antecedent for the consequent to occur. However, the selection of the not-P card may still be informative regarding the truth of the causal rule. Consider for example the two possible outcomes given the closure of the valve, the absence of water flow or the presence of flow. Assuming the device is working we may expect to observe an absence of water flow when the valve is closed. However, this absence may also be caused if the antecedent of the second conditional is not satisfied, that is if the pipe is blocked. Hence observing the absence of the effect in the absence of the cause is not informative about the truth of the rule. However, imagine instead that we observed water flowing when the valve was closed. This case would appear more informative regarding the question of whether the mechanical device is operating correctly. What our results suggest is that some participants consider only the first outcome. Hence, they regard the not-P card as uninformative and do not choose it.

Different Types of Causal Hypothesis?

Although our predictions were induced from the literature on mechanical reasoning, our results are of obvious relevance to the literature on causal cognition. For example, it is interesting to compare our results to those of Green & Over (2000) who examined decision theoretic effects in how people test causal conditional hypotheses such as the following:

(3) If you drink from the well then you will get cholera

Across all of their conditions, the rate of antecedent selections never exceeded the rate of consequent selections to the same degree as was true of our one rule condition. Collapsed across conditions, their rate of consequent selections was, in fact, marginally greater than the rate of antecedent selections (a total of 171 consequent selections vs. 169 antecedent selections). These results are in stark contrast to our own findings.

Green and Over's experiment was designed to test ideas concerning the relationship between the contingency table and causal hypothesis testing. Their results show that people are sensitive to the probabilities of the cause and the effect when deciding which cards to select. We believe that our results are different to theirs because we asked our participants to test a causal *mechanical* rule whereas their experiment concerned a causal *medical* rule. O'Brien, Costa & Overton (1986) have also found results suggesting that there are domain-specific differences in causal reasoning. In their experiment participants were asked what implications each of the four cases (*cause present + effect present*; *cause present + effect absent*; *cause absent + effect present*; *cause absent + effect absent*) had for a variety of hypotheses concerning medical and mechanical causal relationships. For all cases except the *cause absent + effect absent* case, participants were significantly less certain about the medical hypothesis than about the mechanical hypothesis.

One way to conceive of O'Brien et al's result is that people are unwilling to accept a medical hypothesis in the light of information about just a few exemplars whereas they have more confidence about the status of a mechanical hypothesis given a few confirming or disconfirming cases. In other words, causal medical

hypotheses require reasoning that is likely to be probabilistic in nature whereas causal mechanical hypotheses need not (of course it is possible to design a mechanical task that encourages probabilistic reasoning - see Kirby, 1994).

There are several factors that might be involved in causing one type of reasoning to be essentially probabilistic and the other deterministic. First, it is possible that our knowledge about organisms tells us that even in the presence of a cause, the effect might not occur. In other words, causal rules about organisms admit of many disabling conditions. In addition, illnesses have many possible causes. In Green & Over's example, someone drinking from the well might be immune to cholera or cholera might be present in the well, the local river and the nearby lake. Given all of these possibilities it makes sense that participants in O'Brien et al's study were unwilling to make decisions about the status of a medical rule in the light of information about individual exemplars. Similarly, it is unsurprising that participants in Green and Over's experiment evidenced probabilistic thinking.

Now think about testing a mechanical hypothesis. Such hypotheses are normally tested via intervention. That is, if you think that your car won't start because the plugs are dirty, you will clean the plugs and then try to start the car. If the car starts your hypothesis has been confirmed, if not then it has been disconfirmed. In either case it is unlikely that you will repeat the procedure several times in case of the operation of disabling conditions or alternative causes. Similarly, when interacting with a novel electronic device (see Klahr & Dunbar, 1988) people do not perform the same test several times in order to establish the effect of some manipulation. Their reasoning in such cases tends to be non-probabilistic.

This distinction, between probabilistic causal reasoning and consequential (or deductive) causal reasoning also relates to the literature on mechanical reasoning described in the introduction. The systems that Hegarty (1992) and Schwartz & Black (1996) required their participants to reason about were closed and so did not admit of disabling conditions or alternative causes. Of course, a particular manipulation to the system might not cause the expected effect but given the diagrams that people were shown, disabling conditions and alternative causes were unlikely to be available to reasoners. Accordingly, a strategy based on the piecemeal animation of the device in the causally appropriate direction will be adopted. For analogous reasons, our participants were interested in tests of the hypothesis about the cooling system that involved direct manipulations to the putative cause. Once the possibility of disabling conditions were introduced, they were significantly less interested in such tests.

References

Byrne, R.M.J. (1989). Suppressing valid inferences with conditionals. Cognition, 31, 61-83.

Cummins, D.D., Lubart, T., Alksnis, O. & Rist, R. (1991). Conditional reasoning and causation. Memory & Cognition, 19, 274-282.

Feeney, A. & Handley, S. (2000). The suppression of q card selections: Evidence for deduction in Wason's selection task. Quarterly Journal of Experimental Psychology, 53A, 1224-1242.

Green, D.W. & Over, D.E. (2000). Decision theoretic effects in testing a causal conditional. Current Psychology of Cognition, 19, 51-68.

Handley, S.J., Feeney, A. & Harper, C. (2000). Alternative antecedents and the suppression of fallacies in Wason's selection task. Manuscript under review.

Harris, P.L., German, T. & Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning. Cognition, 61, 233-259.

Hegarty, M. (1992). Mental animation: Inferring motion from static displays of mechanical systems. Journal of Experimental Psychology: Learning, Memory & Cognition, 18, 1084-1102.

Johnson-Laird, P.N. (1993). Human and machine thinking. Hove, UK: LEA.

Kirby, K.N. (1994). Probabilities and utilities of fictional outcomes in Wason's four card selection task. Cognition, 51, 1-28.

Klahr, D. &Dunbar, K. (1988). Dual space search during scientific reasoning. Cognitive Science, 12, 1-55.

Mackie, J.L. (1974). The cement of the universe: A study of causation. Oxford, UK: Oxford University Press.

Medvedev, G. (1991). The truth about Chernobyl. London: I.B. Tauris & Co. Ltd.

O'Brien, D.P., Costa, G. & Overton, W.F. (1986). Evaluations of causal and conditional hypotheses. Quarterly Journal of Experimental Psychology, 38A, 493-512.

Popper, K. (1959). The logic of scientific discovery. London: Hutchinson.

Rumain, B., Connell, J., & Braine, M.D.S. (1983) Conversational comprehension processes are responsible for reasoning fallacies in children as well as adults: IF is not the biconditional. Developmental Psychology, 19, 471-481.

Schwartz, D.L. & Black, J.B. (1996). Shuttling between depictive models and abstract rules: Induction and fallback. Cognitive Science, 20, 457-497.

Wason, P.C. (1968). Reasoning about a rule. Quarterly Journal of Experimental Psychology, 20, 273-281.

Acknowledgements

We would like to thank Lara Webber for her assistance in carrying out the experiment described in this paper.

An Influence of Spatial Language on Recognition Memory for Spatial Scenes

Michele I. Feist (m-feist@northwestern.edu)
Department of Psychology, Northwestern University
2029 Sheridan Road, Evanston, IL 60208 USA

Dedre Gentner (gentner@northwestern.edu)
Department of Psychology, Northwestern University
2029 Sheridan Road, Evanston, IL 60208 USA

Abstract

Whether and how much the routine use of language influences thought is a perennially fascinating question in cognitive science. The current paper addresses this issue by examining whether the presence of spatial language influences the encoding and memory of simple pictures.

Introduction

In the last few years there has been a resurgence of interest in the question of whether and how much language influences thought. As Billman and Krych (1998) point out, this is a question that can be asked either at the level of the language system, or at the level of the linguistic form.

At the level of the language system, one can ask whether cognitive differences can be explained via cross-linguistic differences. The strong version of this hypothesis is well expressed in Whorf's (1956, p. 134) quote of Sapir: "[w]e see and hear and otherwise experience very largely as we do because the language habits of our community predispose certain choices of interpretation." Other scholars suggest a weaker version of the hypothesis, namely that language, while not determining thought, nonetheless influences how one thinks. Slobin's (1996) *thinking-for-speaking* hypothesis states that linguistic influences exist only when one performs a linguistically-mediated task (cf., Slobin, 1996).

Evaluation of the hypothesis at the level of the language system involves an examination of performance on non-linguistic tasks by speakers of different languages in order to determine whether there are language-related differences. Such examinations have yielded mixed results. Pederson and his colleagues (1998) and Levinson (1996) found that speakers of different languages performed differently on nonlinguistic tests of visual memory, including reconstruction of an array of objects, a clearly Whorfian result. Malt, Sloman, and Gennari (in press), on the other hand, found that Spanish speakers' judgments of similarity of videotaped motion events conformed to normal verb use in Spanish, but only when participants were instructed to use linguistic descriptions during the encoding phase of the experiment. This is consistent with a thinking-for-speaking (Slobin 1996) version of the Sapir-Whorf hypothesis. Furthermore, the language effect did not appear for the English-speaking participants, nor did Malt and her colleagues find a language effect on similarity judgments for artifacts, nor on recognition memory.

The other level at which language could influence thought is that of linguistic forms within a language. Evaluation of the hypothesis at this level involves comparing performance on non-linguistic tasks by speakers of the same language in conditions that invite different forms within the language. For example, Bower, Karlin, and Dueck (1975) found that participants rated new pictures as more similar to the one they had seen during encoding if they conformed to the linguistic description presented at encoding. Gentner and Loftus (1979) found an influence of the language presented at encoding on participants' recognition memory for pictures of events. Billman and Krych (1998) found effects of verbs present at encoding on recognition of videotaped motion events (but see Malt et al., in press).

Our research asks whether spatial prepositions can influence the way people encode and remember spatial relations. We chose spatial prepositions for several reasons. First, while many studies of the Whorfian question have focused on possible effects of verbs of motion on the encoding of events, there has been comparatively little work on the possible effects of prepositions on the encoding of static spatial relations. Spatial prepositions exhibit striking cross-linguistic variability, as demonstrated by Bowerman and Pederson's (in preparation) comparative study of the semantics of '*on*-terms' – terms related to contact and support. As Gentner (1981; Gentner & Boroditsky, 2001) points out, relational terms such as verbs and prepositions are a promising arena in which to seek Whorfian evidence. Relational terms are more variable cross-linguistically than nominal terms of comparable concreteness. This semantic variability suggests that there is a wide variety of plausible encodings consistent with the perceptual input. Thus, this arena may provide fruitful ground for the investigation of Whorfian effects.

In this research, we showed people spatial scenes under different linguistic encoding conditions, and later tested their recognition memory. Our goal was to determine (1) whether

spatial language influences spatial encoding and memory and (2) whether such influence occurs when there is no overt use of language, or is restricted to the case when spatial language is explicitly present. If we see language effects only when people are encouraged to utilize language at encoding, this will provide support for a thinking-for-speaking or, in our case, *thinking-for-comprehending* hypothesis. If, on the other hand, we see language effects under other conditions, this would leave open the possibility of language influencing cognition in a more comprehensive manner.

The logic of our studies is as follows. For each of the prepositions, we created a sentence and a triad of pictures that ranged in how well they fit the sentence (see Figure 1). The standard picture (the *initial* picture) was acceptably described. For each standard, there were two variants: the *plus* variant, which was a better exemplar of the spatial term, and the *minus* variant, which was a poorer exemplar (see Figure 1 below). Thus, the initial picture was somewhat ambiguous, but was designed so that the spatial term could apply to it, and the two variants were either more typical of the core prepositional category or less so. All of the pictures involved the same objects; the only source of variation was the spatial relation between the two objects. In preparing the pictures, every attempt was made to guard against a possible recognition bias for the *plus* variant (see Experiment 2).

Experiment 1a

Participants viewed pictures depicting static spatial relations - e.g., a marionette standing on a table or a coin in a hand. Half the participants read a descriptive sentence at the time that the pictures were encoded. After participating in unrelated experiments for about fifteen minutes, participants performed a recognition task that included the original pictures and two variants.

The recognition test included all three pictures - the initial picture, the plus variant, and the minus variant. If the presentation of language at encoding influences recognition memory, there should be different patterns of false alarms for the two groups. The group provided with sentences at encoding should be more likely than the control group to falsely claim that they had previously seen the plus variants of the pictures.

Method

Design. Encoding Condition (Spatial Sentences/Control), a between-subjects variable, was crossed with Recognition Item Type (Plus Variant/Initial Picture/Minus Variant), a within-subject factor.

Subjects. Thirty-six Northwestern undergraduates received course credit for their participation in this experiment. All reported being fluent speakers of English.

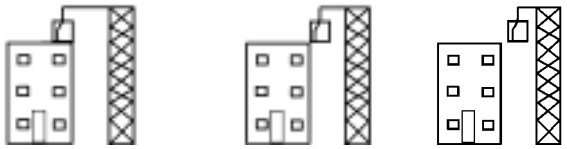
Stimuli. Thirteen triads of pictures and corresponding sets of sentences were created for this experiment. As discussed above, the pictures were created such that one might be well described by a target sentence, one passably described, and one poorly described. Each triad of pictures was associated with a pair of sentences: the target sentence that described the picture as outlined above, and a distracter sentence in which only the nouns were changed. The distracter sentence was meant to be obviously wrong; its purpose was simply to force participants to read the correct sentence and encode the target spatial relational term. For example, for the picture in Figure 1, participants chose between *The block is on the building* and *The plant is on the shelf*.

The initial picture from each triad was used for the study portion of the experiment; all three pictures in the triad were used for the recognition task.

Procedure

Part 1: Study. Twenty-five pictures (thirteen targets and twelve distracters) were randomized and presented individually for five seconds each on a computer screen. All participants were told that this was part one of a two-part experiment.

To ensure that the spatial sentences group processed the sentences we asked them to choose which of two sentences best described the picture. They were provided with answer sheets with two sentences for each picture: the target sentence and a distracter sentence. Participants in the control
condition were given no additional instructions.



Plus variant *Initial picture* *Minus variant*

Figure 1: Triad of pictures corresponding to the sentence "The block is on the building."

Part 2: Recognition. All participants received the same yes/no recognition task. All three of the pictures in each triad were presented individually in random order along with twelve distracters (six old and six new). Participants were asked to indicate on a numbered answer sheet whether or not they had seen each picture during the earlier study portion. Each picture remained on the screen until the participant pressed the "c" key, indicating that they were ready to continue.

Results

As predicted, we found that participants' recognition memory was influenced by whether a linguistic description was presented during study. Participants in the spatial sentences condition were significantly more likely to false-alarm to the *plus* variant than to the *minus* variant. (Figure 2). The difference between the false alarms in response to the *plus* variant and the false alarms in response to the *minus* variant differs significantly in the spatial sentences condition, as confirmed by a paired samples t-test ($t(17) = 5.32$, $p < .0001$). Participants in the control condition showed no such difference in their false alarm rate. Thus,

having spatial language present at encoding led to a skewing of recogntion errors towards the core of the spatial category.

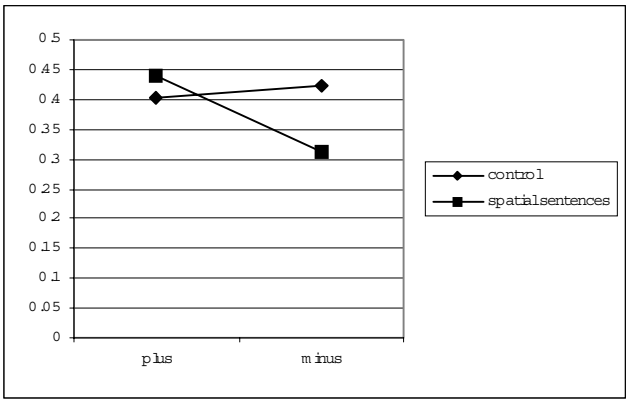


Figure 2: False alarms by condition, Experiment 1a

d' analysis To further test the claim that the presentation of sentences during study influences recognition memory for pictures, two d' measures were calculated for each individual subject. One d' indicates the discriminability of the *minus* variant and the initial picture; the other, the discriminability of the *plus* variant and the initial picture. The larger of the two was then determined, and the participants were pooled by condition, as shown in Table 1.

Table 1: Participants pooled according to the d' analysis, Experiment 1a

	Plus larger	*Minus* larger	Equal[1]
Control	4	4	10
Spatial Sentences	0	12	5

In the spatial sentences condition, but not in the control condition, the discriminability of the *minus* variant is greater than that of the *plus* variant (X^2=9.65, p<.01).

Discussion

We found that when spatial language was present at encoding, memory for the spatial relations in the pictures was systematically shifted in the direction of the spatial preposition. This is evidence for at least the moderate thinking-for- speaking version of the Whorfian hypothesis. In the next study we sought evidence for the strong version of the hypothesis. We hypothesized that if people had to attend closely to the pictures, this might evoke spontaneous linguistic descriptions as a memory aid. We thus examine the effect of more careful attention on recognition memory in Experiment 1b.

[1] d' measures within .25 of one another were considered equal for the analyses discussed in this paper.

Experiment 1b

In this study we asked whether participants instructed to pay careful attention to the pictures at study might be induced to encode the pictures linguistically and, as a result, to display an error pattern similar to that seen in the spatial sentences condition of Experiment 1a.

Method

Subjects Eighteen Northwestern undergraduates received course credit for their participation in this experiment. All reported being fluent speakers of English.

Stimuli The stimuli used were the same as those in Experiment 1a.

Procedure

Part 1: Study The procedure was identical to the control condition in Experiment 1a, except that the participants were instructed to pay careful attention to the pictures because the recognition test would be very difficult.

Part 2: Recognition The recognition task was the same as that used in Experiment 1a.

Results and Discussion

The error rate observed in Experiment 1b is lower than that observed in Experiment 1a, indicating that participants did pay more careful attention to the pictures during study. However, the pattern of false alarms is the same as that observed for the control subjects from Experiment 1a. Figure 3 shows the results of Experiment 1b along with those of Experiment 1a. These results suggest that more careful attention did not necessarily evoke linguistic encoding.

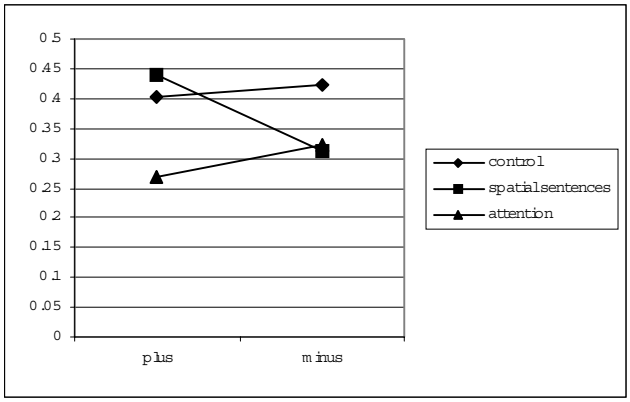


Figure 3: False alarms by condition, Experiments 1a and 1b

So far we have evidence for the influence of spatial language when it is explicitly presented, although not for the stronger possibility that language will affect cognition even when it is not overtly present. In Experiment 1c, we tested the specificity of the language effect. If, as we have assumed, the recognition shift is due to spatial language,

then we should not see this shift if participants are given verbal descriptions that do not contain spatial language.

Experiment 1c

In order to more carefully inspect the source of the language effect from Experiment 1a, we presented participants with sentences without spatial prepositions at encoding. The sentences used named only the objects in the picture. We predict that these sentences, which are missing the hypothesized source of the language effect, will not replicate the effect found in Experiment 1a.

Method

Subjects Nineteen Northwestern undergraduates received course credit for their participation in this experiment. All reported being fluent speakers of English.

Stimuli The pictures were the same as those in Experiment 1a. The sentences on participants' answer sheets were modified from those used in Experiment 1a by removing the prepositions, resulting in sentences of the following form:

The picture shows a block and a building.
The picture shows a plant and a shelf.

Procedure

Part 1: Study The procedure was identical to that in the spatial sentences condition in Experiment 1a. Participants chose which sentence best matched the picture.

Part 2: Recognition The recognition task was the same as that used in Experiment 1a.

Results and Discussion

As predicted, participants failed to show any shift towards the core spatial category designated by the preposition. The participants in Experiment 1c demonstrated the same pattern of equal *plus* and *minus* false alarms as the no-language subjects in the previous studies (the subjects in Experiment 1b and the control subjects in Experiment 1a). This pattern differed significantly from the pattern by spatial sentence subjects in Experiment 1a. Specifically, the two groups differed in their rate of false alarms in response to the minus variant (independent samples t-test: $t(34) = 3.91$, $p<.005$). This provides support for the suggestion that it is specifically the preposition that is responsible for the change in the pattern of responses observed in the spatial sentences condition in Experiment 1a. The complete set of results for Experiment 1 is presented in Figure 4.

d' analysis As in Experiment 1a, two d' measures were calculated for each individual participant in Experiment 1: one indicated the discriminability of the *minus* variant and the initial picture, and one indicated the discriminability of the *plus* variant and the initial picture. The larger of the two was then determined, and the participants were pooled by condition (Table 2).

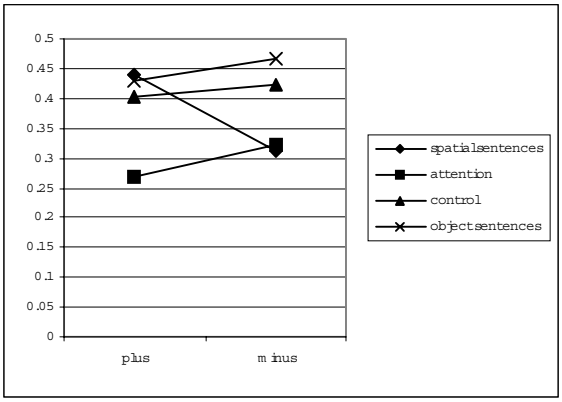


Figure 4: False alarms by condition, Experiment 1

Table 2: Participants pooled according to the d' analysis, Experiment 1

	Plus larger	*Minus* larger	Equal
Control	4	4	10
Spatial sentences	0	12	5
Attention	8	4	5
Object sentences	6	3	7

In the spatial sentences condition only, the discriminability of the *minus* variant is greater than that of the *plus* variant ($X^2=19.31$, $p<.01$). Or to put it more directly, only in the spatial sentences condition is the *plus* version more confusable with the initial picture than the *minus* version.

Experiment 2

This study was done to verify that the spatial sentences applied to the three variants of each picture as expected. We asked participants to rate the applicability of the sentences from the study portion of Experiment 1a to each of the pictures.

Method

Subjects Twenty-four Northwestern undergraduates received course credit for their participation in this experiment. All reported being fluent speakers of English.

Stimuli The pictures used were the same as those in Experiment 1. The sentences used were the correct spatial sentences from Experiment 1a.

Procedure

All three of the pictures in each triad were presented individually in random order along with the twelve distracters from the recognition task from Experiment 1. Participants were asked to rate the applicability of the sentences to the pictures on a scale from one to seven, with seven being the highest rating. Each picture remained on the screen until the participant pressed the "c" key, indicating that they were ready to continue.

Results and Discussion

As expected, participants gave the highest ratings to the *plus* variants (mean rating 5.72), in-between ratings to the initial pictures (mean rating 4.47), and the lowest ratings to the *minus* variants (2.54). This distribution of the ratings suggests that the assignment of pictures to the various categories with respect to the sentences used in the spatial sentences condition of Experiment 1a was indeed appropriate. Examination of the results for individual triads showed that for two of the triads, one depicting a coin in a hand and one depicting a firefly in a dish, the sentences did not fit exactly as predicted. These sentences were adjusted accordingly for Experiment 3.

Experiment 3

This study was a replication of the spatial language condition, with a methodological improvement. In Experiment 1a, participants saw all three versions of each of the pictures (one at a time) during the yes/no recognition task. This leaves open the possibility of carryover effects from one variant to another. In Experiment 3, the study task was that of Experiment 1a, but the recognition task was designed so that each participant was tested on only one version of each picture.

Method

Design. Encoding Condition (Spatial Sentences/Control), a between-subjects variable, was crossed with Recognition Item Type (Plus Variant/Initial Picture/Minus Variant) (within-subjects) and with Assignment condition. This was a between-subjects variable determining which variant in each set was received by a given participant in the recognition test.

Subjects. One hundred eighteen Northwestern undergraduates received course credit for their participation in this experiment. All reported being fluent speakers of English.

Stimuli. The stimuli used were the same as those in Experiment 1, with minor modifications to two of the triads of pictures, and with a change of preposition (from *in* to *on*) in the sentences corresponding to two others. One of the triads used in Experiment 1, depicting a balloon on a stick, was not used for Experiment 3.

Procedure

Part 1: Study The procedure was identical to the study portion of Experiment 1a.

Part 2: Recognition Both conditions received the same yes/no recognition task. One picture from each triad was presented in random order along with twelve distracters (six old and six new). As in Experiment 1, participants were asked to indicate whether or not they had seen each picture during the earlier study portion, and each picture remained on the screen until the participant pressed the "c" key indicating readiness to continue.

Results

As in Experiment 1a, we found that participants' recognition memory was influenced by the presence or absence of spatial language during study. The pattern of false alarms for the spatial sentences condition differs from that in the control condition (Figure 5). As in Experiment 1a, participants in the spatial sentences condition were significantly more likely to false-alarm to the *plus* variant than to the *minus* variant. Participants in the control condition showed no such difference in their false alarm rate. The difference between the false alarms in response to the *plus* variant and the false alarms in response to the *minus* variant differs significantly only in the spatial sentences condition, as confirmed by a paired samples t-test ($t(57) = 2.23$, $p = .047$). In addition, the difference in the rate of false alarms between the two groups only reaches significance for the responses to the *plus* variant, as confirmed by an independent samples t-test ($t(116) = 2.20$, $p = .039$).

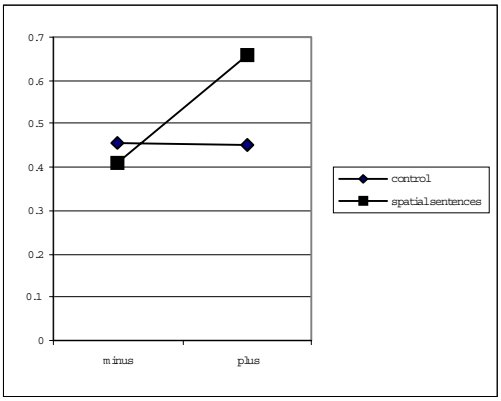


Figure 5: False alarms by condition, Experiment 3

d' analysis As in Experiment 1a, two d' measures were calculated for each individual subject. One d' indicates the discriminability of the *minus* variant and the initial picture; the other, the discriminability of the *plus* variant and the initial picture. The larger of the two was then determined, and the participants were pooled by condition (Table 3).

Table 3: Participants pooled according to the d' analysis, Experiment 3

	Plus larger	*Minus* larger	Equal
Spatial sentences	4	38	16
Control	20	20	20

The results of the d' analysis for Experiment 3 replicate those for Experiment 1: in the spatial sentences condition alone, the discriminability of the *minus* variant is greater than that of the *plus* variant ($X^2 = 16.67$, $p < .0001$).

General Discussion

In these experiments, we examined the question of whether spatial language influences the encoding and memory of spatial relations presented visually. The answer is a qualified yes. Our evidence shows that the use of spatial language during the encoding of a picture can affect recognition memory for the spatial relations in the picture. People given spatial prepositions during encoding showed a shift in recognition towards the core spatial category denoted by the preposition (Experiments 1a and 3). This effect was specific to spatial relational language (Experiment 1c); no such shift was observed for sentences that simply described the objects in the pictures.

However, our evidence that language influenced encoding was limited to the case when overt spatial language was present. We did not find a shift towards the core spatial semantic category when participants were simply instructed to pay close attention to the pictures (Experiment 1b). Thus, our evidence supports the view that language can affect encoding when it is present, but not the strong Whorfian view that non-linguistic perception is shaped by the language one speaks.

There has been much controversy in recent years over whether language exerts an effect on non-linguistic cognition. Our results suggest that language forms do exert an effect on one type of non-linguistic cognition: recognition memory for simple pictures. This suggestion must be qualified, however, as we do not show an effect of language forms in the absence of linguistic descriptions at encoding, which would suggest a stronger influence of language on everyday non-linguistic cognition. Of course, it remains an open question whether in some situations, speakers might prefer encodings that are compatible with their language, resulting in cross-linguistic differences that are habitual though not inescapable.

Our results are compatible with Slobin's (1996) thinking-for-speaking hypothesis and with the results of Malt et al. (in press). They suggest that language can have profound non-linguistic effects when it is used, but that its use is not inevitable. This is consistent with Gentner and Loewenstein's (in press) suggestion that language provides tools that potentiate forming and holding ideas -- the *tools–for-thought* hypothesis. On this view, language potentiates kinds of encodings rather than forcing them.

Acknowledgments

Please address all correspondence and reprint requests to Dedre Gentner, Northwestern University, Department of Psychology, 2029 Sheridan Road, Evanston, IL 60208. This work was supported by NSF-LIS grant SBR-9720313 to the second author.

References

Billman, D., & Krych, M. (1998). Path and manner verbs in action: Effects of "skipping" and "exiting" on event memory. *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Bower, G. H., Karlin, M. B., and Dueck, A. (1975). Comprehension and memory for pictures. *Memory and Cognition, 3 (2), 216-220.*

Bowerman, M., and Pederson, E. (in preparation). Cross-linguistic perspectives on topological spatial relationships.

Gentner, D. (in press). Why we're so smart. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought*. Cambridge, MA: MIT Press.

Gentner, D. (1981). Some interesting differences between verbs and nouns. *Cognition and Brain Theory,4 (2)*, 161-178.

Gentner, D., & Boroditsky, L. (2001). Individuation, relativity and early word learning. In M. Bowerman & S. Levinson (Eds.), *Language acquisition and conceptual deveolopment*. Cambridge, England: Cambridge University Press.

Gentner, D., & Loewenstein, J. (in press). Relational language and relational thought. In J. Byrnes & E. Amsel (Eds.), *Language, literacy, and cognitive development.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gentner, D., & Loftus, E. (1979). Integration of verbal and visual information as evidenced by distortions in picture memory. *American Journal of Psychology, 92 (2),* 363-375.

Levinson, S.C. (1996). "Relativity in spatial conception and description." In Gumperz, J. and Levinson, S. (Eds.), *Rethinking Linguistic Relativity*. Cambridge, England: Cambridge University Press.

Malt, B. C., Sloman, S. A., & Gennari, S. (in press). Speaking vs. thinking about objects and actions. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought*. Cambridge, MA: MIT Press.

Pederson, E., Danziger, E., Wilkins, D., Levinson, S. C., Kita, S., & Senft, G. (1998). Semantic typology and spatial conceptualization. *Language, 74 (3),* 557-589.

Slobin, D. (1996). From "thought and language" to "thinking for speaking." In J. J. Gumperz and S. C. Levinson (Eds.), *Rethinking linguistic relativity.* Cambridge: Cambridge University Press.

Whorf, B. L. (1956). *Language, thought, and reality: Selected writings of Benjamin Lee Whorf.* J. B. Carroll (Ed). Cambridge, MA: MIT Press.

The Origin of Somatic Markers:
A Suggestion to Damasio's Theory Inspired by Dewey's *Ethics*

Suzanne Filipic (suzannefilipic@hotmail.com)
Université de Paris III-Sorbonne Nouvelle
39 avenue de la citadelle 21240 Talant France

Damasio vs. Dewey

My goal is not to propose a criticism of Damasio's theory, but to suggest how it might be possible to carry it further, to wipe out even more efficiently the traditional dualism opposing our reason and emotions. Damasio's theory of somatic markers is for me a very efficient means of exposing the importance of emotions on the workings of reason, but my fear is that it might lead a few to replace the old dualism opposing an efficient reason to disruptive emotions by a new one overemphasizing the power of emotions on a weak influenced reason.

An important part of Dewey's *Ethics* deals with this dualism. I suggest that, although Damasio's experiments can be seen as providing amazing clarity and precision to Dewey's philosophical intuitions, Dewey's elaboration on the reciprocal influence of reason and emotions goes one step further in questioning whatever supposedly opposes them.

Dewey's *Ethics*

Since Dewey's *Ethics* is probably much less known today than Damasio's *Descartes' Error*, let me start by presenting what interactions Dewey imagines between our emotions and our reason.

For Dewey, when we think we are making a choice between following our emotions, or following our reason, reality is always more complex:

> while there is conflict, it is not between desire and reason, but between a desire which wants a near-by object and a desire which wants an object which is seen by thought to occur in consequence of an intervening series of conditions, or in the "long run". (1932)

When facing a new situation, when are emotions necessary if thought must lead us to action ? Dewey's first intuition is that emotions provide the necessary starting point of reflection, or the energy necessary to its activity:

> Unless there is a direct, mainly unreflective appreciation of persons and deeds, the data for subsequent thought will be lacking or distorted. A person must feel the qualities of acts as one feels with the hands the qualities of roughness and smoothness in objects, before he has an inducement to deliberate or material with which to deliberate. (1932)

Dewey also thinks that any outcome of a thinking process must be emotionally "appreciated," otherwise it would not stir us to action. I quote *Ethics* again: "no matter how elaborate and how rational is the object of thought, it is impotent unless it arouses desire".

At this point, we might wonder two things. Are the emotions which stimulate reflection and those which motivate action of the same kind? And doesn't Dewey overemphasize the importance of emotions, by giving them a crucial role at the beginning and at the end of the process of thought?

In the fourteenth chapter of *Ethics*, Dewey works on distinguishing valuation, as a judgment of value, in which reason evaluates an object by its consequences, and valuing, as an immediate emotional reaction. For Dewey, valuation and valuing are not opposed, but linked: "We esteem before we estimate, and estimation comes in to consider whether and to what extent something is worthy of esteem [...]. All growth in maturity is attended with this change from a spontaneous to a reflective and critical attitude". (1932)

Therefore, even if emotional reactions always come first, reason can and should have an effect upon later ones: "judgments of value are not mere registrations [...] of previous attitudes of favor and disfavor, liking and aversion, but have a reconstructive and transforming effect upon them by determining the objects that are worthy of esteem and approbation". (1932)

Dewey thus distinguishes between primary emotions, which one has at the beginning of one's life, and adult emotions. These ones are of two different kinds: the spontaneous ones, which are immediate, but probably the result of past value judgments, and the transformed emotions, which have just evolved, as an effect of a new value-judgment.

Therefore, to go back to my two questions, we can conclude that Dewey ends up giving emotions an important influence on reason, and the two kinds of emotions he imagines in this regard might coincide with the two kinds of emotions he had noticed as necessary to reflection. The emotions providing energy to the thought process might be the spontaneous ones, occurring before the thought process, and the transformed emotions, produced by valuation, might be the ones necessary to act.

If one accepts this interpretation, we could conclude that in front of a situation, Dewey imagines that our reaction follows a pattern similar to this one: a new situation *provokes* spontaneous emotions *which stimulate* reflection *which produces* a value-judgment and a transformed emotion *which enable us to act*.

What suggestion will this theory enable us to make on Damasio's theory of somatic markers?

Damasio's Theory of the Somatic Markers

The starting point of Damasio's research is Elliot, a reasonable and intelligent man, who, because of a brain tumor, became unable to take any sound personal decision.

After running a series of tests, Damasio became convinced that Elliot was still able, in front of most situations, to imagine different action plans, but that he was never able to choose the right one in practice. To link his lack of emotions to his inability to assign values to the different plans he is still able to produce, Damasio proposes his theory of the somatic marker.

To explains what he means by a "somatic marker," Damasio asks his reader to imagine himself as an owner of a large business, "faced with the prospect of meeting or not with a possible client who can bring valuable business but also happens to be the archenemy of your best friend, and proceeding or not with a particular deal" (1994).

For Damasio, using a cost/benefit analysis of all the scenarios you imagine is not going to work; at best, it would take you too long to make a decision. However, he thinks that without reasoning about it, some of the options you imagine are automatically eliminated. If, from experience, a connection has been made between a specific response option and its bad outcome, a somatic marker will be activated. This marker would then operate either outside consciousness, by inhibiting a tendency to act, or consciously, by letting you experience an unpleasant gut feeling, thus convincing you to avoid this option.

Damasio's Theory as an Explanation of Dewey's Intuitions

Before going any further, I would like to suggest that Damasio's theory can first be read as an explanation of Dewey's relatively vague notions. When Damasio writes, " a somatic state, negative or positive, caused by the appearance of a given representation, operates not only as a *marker for the value of what is represented, but also as a booster for continued working memory and attention"* (1994), it can easily be read as an explanation of what Dewey meant when he wrote that emotions provide reflection with "material with which to deliberate," and "an inducement to deliberate" (1932).

The Origin and Evolution of Somatic Markers

Somatic markers are, according to Damasio, what enable us, in new circumstances, to experience feelings before we start evaluating the situation rationally. For Dewey, these "intuitions" which are necessary for reflection because they provide its material and its motivation, are in the long run the product of our value-judgments. Thus, he estimates that our emotions are as necessary to our reason, than our reason to our emotions.

I think that Damasio considers that only emotions are at the source of somatic markers. Since he sees them as necessary to the thinking process, it leads him to conclude that our reason is based on our emotions, but he forgets to consider whether our emotions are influenced by our reason.

At this point, my goal is to question Damasio's apparent one-sided view on the reason/emotions interaction. My first remark is simply theoretical. Following Damasio's experiments, I think we can reach different conclusions. For him, when a chosen option leads to a negative outcome, the consecutive somatic state (the painful emotion) allows a new marker to be created. Damasio thus insists on the fact that emotions are what create somatic markers.

However, I think it is just as logical to conclude that, since we also use our reason to choose an option, our reason is also an important cause of the resulting somatic state, and thus of the new marker. In other words, if our reason had enabled us to make a better choice, the final emotion would have been different, and therefore the new marker would have been different.

I think Damasio does consider reason as an important step in the decision-making process. Yet, because his experiments led him to re-evaluate the importance of emotions, he stopped at the conclusion that emotions are the foundations of reason, and it might be asked whether we should not also consider that our emotions might be just as much the products of our reason.

My second remark is about Damasio's main experiment. I think that Damasio ends up giving emotions such a one-sided influence on reason because the experiments he works from are cases in which the role of emotions is much more important than any conscious evaluation, in the production and the evolution of somatic markers. I will try to analyze the "gambling experiments" proposed in chapter 9 of *Descartes' Error.*

I think it may be necessary to distinguish between two different processes: the production, and the evolution of somatic markers. My point is not to suggest that these two processes are totally distinct, but

rather that the second one progressively distinguishes itself from the first. If these two processes are thought of as different, I think we can consider how much Damasio's experiments do reproduce real-life decision-making circumstances, and how much they differ from them.

Damasio's experiments consist of asking his patients and "normal" individuals to gamble, playing with four decks of cards, two decks giving out high rewards but also high penalties, and thus leading the players to bankruptcy, and two other decks, causing lower rewards but also much lower penalties, enabling the players to win the game. These experiments are a success since they enable Damasio to distinguish patients, who loose the game, from normal individuals, who win, because they learn to avoid the bad decks of cards.

If we consider how somatic markers are created, which means if we consider a limited number of the same kind of experiences, I think Damasio's gambling experiments reproduce what happens in real life. An experience which leads you to a success, produces a positive marker which will be activated in the future if the same circumstances are experienced again (in this case, decks C and D); whereas an experience which ends as a failure produces a negative marker (as for decks A and B).

However, if we consider the way if which markers evolve, I think these gambling experiments only allow us to study a limited category of our real-life experiences. If we consider "normal" individuals, after a few cards have been turned down, when the question comes up again to choose a deck of cards, the different stages of the decision-making process are the following. After the different response options have been produced (in this case, there will be four options, since there are four decks of cards), the negative marker associated with two of the four options allows them to be eliminated quickly, and then a choice has to be made between the two remaining ones, either automatically, or consciously. When the decision is made, the player picks up a card, experiences an emotion (positive or negative, depending on the efficiency of the preceding markers), the emotion then makes the marker evolve, reinforcing it, if the bet was successful, or modifying it, if it was not the case.

However, I think we can wonder whether this process does not permit survival only if one repeatedly faces the same sets of circumstances (a hundred cards have to be turned down, for the experiment), in which the possible response options are always exactly the same (make a choice between four decks of cards), and in which response options have very similar consequences. Is it the case in real life? Survival in our societies might require much more complex decision-making processes. Damasio was obviously trying to simplify a typical decision-making experience when he devised these gambling experiments. However, these experiments probably do not enable us to consider the importance of another "typical" decision-making experience, where conscious evaluation has a much more decisive role.

In other words, we might suggest that as long as an environment is stable, human needs do not evolve, and thus the situations or the objects that humans look for are always similar. Their survival is much more easy to achieve if an automatic process (like the action of somatic markers, if we accept Damasio's theory) enables them to predict the outcome of familiar experiences. Yet, in a constantly evolving environment, in which many experiences are unique, an automatic decision-making process might not always be the most efficient one.

Another difference seems essential between Damasio's experiment and life. Damasio says from the beginning that this game is like life because chance rules it. Then, to explain why this test enables him to measure so well his patients' errors in decision-making, he writes that, like in real-life, this test gives the possibility to make choices, but the player does not know neither how, nor when, nor what to choose.

These two passages are for me very surprising, and I think Damasio would agree with me that an individual successful "at the game of life" does not always make a good poker player, and vice versa. Why? Because successes and failures in life usually have a cause, whereas in the experiment they do not. Successes and failures in life can be analyzed, whereas the rules of the gambling experiment, because they are arbitrary, by nature resist analysis. When going through Damasio's experiment, it is necessary to choose the "wrong" decks several times before being cautious because nothing can explain that choosing a particular deck will be, on the whole, a bad option. The only way to persuade oneself is to repeat the mistake.

In life, failures probably encourage analysis a lot more easily. It is not necessary to burn oneself many times to be cautious with fire or hot objects. The first time a child burns himself, he can learn only to never touch again the same kind of object. However, the second time, he has to wonder what it is that these two objects have in common, that makes them objects to avoid.

What is it that enables us to learn from experience, in all experiences where chance is not the strongest element? Our ability to compare experiences, to analyze them, to deduce rules of behavior from individual occurrences, in a word, our reason, even if it is motivated by somatic markers.

To summarize my position, I would say that these gambling experiments are a success because they are an efficient test to distinguish normal individuals from patients. Moreover, skin conductance tests show that it

is probably because a somatic state is activated in normal individuals before they make a decision, that their actions are beneficial on the long term. These experiences thus verify that the activity of somatic markers is a necessary condition if decision-making processes are to help the organism survive. The patients do not succeed at this game, as they do not succeed "in life," because they are unable to produce new somatic markers.

However, these games do not prove that emotions are overall a more important factor than reason in influencing the evolution of somatic markers. The patients might loose the game because their emotions do not produce markers, but they might make wrong decisions in life because their emotions and value-judgments combined do not produce somatic markers either.

The Limited Value of Intuitive Appraisals

Decisions made automatically or unconsciously, on which reason does not have any influence, appear to me as of a limited value, to repeat Dewey's words.

> There is a permanent limit to the value of even the best of the intuitive appraisals [...]. These are dependable in the degree in which conditions and objects of esteem are fairly uniform and recurrent. They do not work with equal sureness in the cases in which the new and unfamiliar enters in. (1932)

The mechanism Damasio describes is probably the one which is at work in his experiments, and in all real-life experiences which have to be undergone in order to survive. (Or at least his book convinced me that this was the case). In these experiences, an automatic learning process can take place, and this mechanism probably enables us to avoid the mistakes we already made.

But what are the processes that enable us to make decisions, when survival is secured? What is the role of conscious reasoning in those processes, probably the last to have appeared in evolution, and still the least important in quantity, that enable us to imagine a solution to a new problem, or a new solution to an old problem, a melody, a new energy?

If I insist on the importance of conscious reasoning on the evolution of somatic markers, it is not to suggest that Damasio does not sufficiently consider the share of conscious reasoning in each experience, but rather to reevaluate the influence of past value-judgments on the unconscious processes that each experience activates.

I do not oppose any element of the somatic marker theory. I only suggest that Damasio might have underevaluated the importance of reason in the long-term evolution of somatic markers, and therefore in our subjective experience, probably because he mostly wanted to demonstrate how limited the influence of reason was on the short-term, when we choose to act in response to a situation.

What are the Consequences of Each Theory?

Damasio's theory can be summarized very briefly as: emotions are what enable us to produce markers, make them evolve, and thus emotions are the necessary conditions of the functioning of our reason.

We can conclude, with Damasio, that emotions have a crucial role, worry that they are given so much importance when their mechanisms are not yet understood: "What worries me is the acceptance of the importance of feelings without any effort to understand their complex biological and sociocultural machinery," and want the fragility of the "foundations" of reason to be recognized:

> The idea of the human organism outlined in this book, and the relation between feelings and reason that emerges from the findings discussed here, do suggest, however, that the strengthening of rationality probably requires that greater consideration be given to the vulnerability of the world within. (1994)

Or, with Dewey, we can estimate that if emotions do have a great influence on the workings of reason, our reason can also influence our emotions. Our rationality is probably fragile, because it is based on emotions, but it is "constructible," since our conscious choices probably have in return a strong influence on the evolution of our emotions.

This process might even be just as automatic as the first one. Our valuations end up modifying our tastes. Without even wanting to change, we do not always like as adults what we liked as children. However, Dewey thinks that some of our "intuitions" (we could probably say: markers) resist analysis:

> The very fact of the early origin and now unconscious quality of the attendant intuitions is often distorting and limiting. It is almost impossible for later reflection to get at and correct that which has become unconsciously a part of the self. (1932)

To conclude on this, I would like to suggest that the task Dewey assigns to reason, that of evaluating if our spontaneous emotions are the result of sensible evaluations, may be a feasible one if one follows the precautionary measure given by Damasio. I suspect that from education we might be able to learn either to mind our somatic states and analyze their causes, or to ignore them.

For example, in the situation imagined by Damasio where you wonder if you should meet a potential client, who happens to be the enemy of your best friend, I think that if one asks himself the question consciously, it may be possible to perceive one's somatic states, and

thus to decide whether to "follow their advice" or not. However, if one pretends to ignore them, instead of diminishing their influence, and let reason work freely--according to those who think that we should not let emotions interfere with reason, their influence will probably be even more important. If an organism learns to ignore its somatic states, the markers will influence the decision-making process anyhow, but without giving reason a chance to influence the decision.

Two Definitions of Reason

In the end, their contrasted definitions of reason seem to be what prevents Dewey and Damasio's theories to coincide. What is reason?

For Damasio, reason seems to be a faculty. When Damasio takes the example of choosing whether to meet your best friend's enemy or not, he opposes his somatic marker hypothesis, to a "pure" reason hypothesis. This is the passage I want to analyze here. For Damasio, pure reasoning will at best enable us to make a decision, but after "an inordinately long time" (1994). However, he thinks that in most cases, a decision will be impossible to make for two reasons.

First, Damasio evokes the limits of our attention and working memory. However, this does not seem to be a sufficient argument, otherwise it would suggest that the patients Damasio works with could solve their problems if they only took a paper and pencil when they need to make a decision. Their cases would probably not have inspired so many ideas to Damasio, if the solution to their problems was that simple.

His second argument is simply that reason's strategies can often be defective. What "strategies" does he have in mind? The answer is for me very surprising, it is the "humans' devastating ignorance and defective use of probability theory and statistics" (1994).

Do we mostly face pure chance? Are probability and statistics calculations our only way to evaluate how others behave, or how society works? Is the reflection on the causes of what happens to us, which should enable us to predict the consequences of what we will choose to do, an impossible task? Because somatic markers enable us to assimilate automatically and to a certain extent the recurrences of reality, can reason only face the "rest," which would be pure chaos?

Shouldn't we consider that after many experiences, we retain not only new somatic markers, which when activated will be able to arouse future somatic states, but that we also retain "markers" of a different kind, which might enable us to make positive choices, and that we call, for lack of more scientific terms, ideas, conscious value criteria?

Why doesn't Damasio write about reason's acquisitions? His theory seems to oppose not only conscious reasoning and the automatic unconscious selection process, but also reason as an empty faculty, and emotions as a content, which can be modified by experience.

Yet, this opposition seems to be more linguistic than proved experimentally. Reason is usually defined (in French as in English) as a capacity, and emotions as states, and it seems that Damasio ratifies this dualism. The fact that his patients are unable to learn from experience may prove that the memory of conscious ideas depends on the activity of somatic markers, but it does not prove that it does not exist in normal individuals.

For Dewey, our reason works from ideas, acquired through one's personal experience and through communication:

> Experience is intellectually cumulative. Out of resembling experiences general ideas develop; through language, instruction, and tradition this gathering together of experiences of value into generalized points of view is extended to take in a whole people and a race. Through intercommunication the experience of the entire human race is to some extent pooled and crystallized in general ideas. These ideas constitute principles. We bring them with us to deliberation on particular situations. (1932)

Just once, Damasio speaks of the necessity of possessing a logical strategy, that would evolve with experience. Was this remark only about our capacity to use statistics and probability better and better?

The primary task of our reason may be to help us reach goals rather than to help us avoid unfavorable situations. It seems probable to consider that in evolution, where one mechanism is sufficient (the somatic markers), a second one does not try to accomplish the same things. At the risk of oversynthetizing Dewey's and Damasio's theories, I think we can suggest that, in case of a failure as of a success, the goal of a thought process will probably be to reach by analysis a plan of future action (better than the one imagined before the just-accomplished action), or even simply to define a set of necessary conditions in hope of attaining this new goal. This would create another kind of "marker," a positive one, which would encourage action, if this set of conditions happens to be experienced in the future. Emotions would then be more efficient at composing a memory of the past, and reason better at building a memory of the future, to quote Damasio's phrase ("memories of the future").

On Strength of Will

Finally, the difference in Dewey's and Damasio's definition of strength of will seems to be very significant of how Dewey considered more than

Damasio the possible consequences of a joint activity of our capacities to reason and to experience emotions.

For Damasio, strength of will is what enables us to endure something painful short term, in exchange for positive consequences on the long term: "Willpower is just another name for the idea of choosing according to long-term outcomes rather than short-term ones" (1994). Willpower can be explained by the action of a positive marker, reason is not evoked.

However, the examples he chooses to illustrate this definition are not decisions one takes easily. On deciding whether to undergo yet another surgery, one might have to decide for it, although it might mean to need to overcome strong negative feelings. The automatic decision-making processes do not seem sufficient in this case.

For Dewey, it is neither reason alone, nor a positive somatic marker, but the product of the union of both, a well thought-of judgment and a "transformed" emotion, that enable us to think in the long term:

> In reality "strength of will" (or, to speak more advisedly, of character) consists of an abiding identification of impulse with thought, in which impulse provides the drive while thought supplies consecutiveness, patience, and persistence, leading to a unified course of conduct. (1932)

Anti-dualism

It might have seemed that I was "defending" the primacy of Reason, but it was not my goal. I think Damasio's theory is essential because it brings to light how necessary emotions are to decision-making processes. I did not intend to refute this, and to argue that reason was more important to decision-making than emotions.

It just seemed appropriate to recall its importance so that, doing away with the dualism opposing an efficient reason to disturbing emotions, we would not clear the way for a new one, opposing influent emotions to an influenced reason. Dewey's hypotheses, as vague and intuitive as they are, seem to sketch a more vague but more global scheme of the reciprocal influence of emotions on reason, and vice versa.

The exception to his theory which Damasio evokes in *The Feeling of What Happens*, the pianist Maria João Pires, who can control by will whether she experiences the emotions that music arouses in her, evokes for me the possibility that we might be considering for now only a very slight portion of the possible interactions between reason, or consciousness, willpower, and emotions.

To come back to my remarks on Damasio's theory suggested by Dewey's writings, I do not know what Damasio would think of them. The theoretic starting point of my analysis, on the "not enough" anti-dualist character of Damasio's theory can seem arbitrary. However, it was precisely the goal of this reflection, to try to show that Dewey's anti-dualism, though a theory, might well be a roundabout way to question reality without being influenced by dualisms handed down to us by culture. Even if we should hope that science will one day have exhausted the hypothesis "resources" of John Dewey's philosophy, I hope I suggested that his anti-dualism can still today inspire scientific research, and thus resolve, if only momentarily, the dualism which so frequently opposes scientific research to philosophical research.

To come back to my fear of seeing a new dualism replace an old one, I will end my discussion by noting that Dewey and Damasio agree in pointing to the dualism opposing mind and body as one of the major sources of (what Stephen Jay Gould calls) "our lamentable tendency to taxonomize complex situations as dichotomies of conflicting opposites" (2000). I think Dewey would have been delighted to hear Damasio correct Descartes' error: "We are, and then we think, and we think only inasmuch as we are, since thinking is indeed caused by the structures and operations of being". (1994)

Acknowledgments

I would like to thank Marie-Christine Lemardeley Cunci (Paris 3) for accepting this research as a possible master's degree dissertation, Christiane Chauviré (Paris 1) for her interest in this novice analysis of Dewey's works, and William Schubert (UIC) for his encouragements. I would also like to thank the French Department of the University of Illinois at Chicago for the 1999-2000 teaching assistantship they awarded me, which enabled me to complete this research.

References

Damasio, A. R. (1994) *Descartes' Error-Emotion, Reason and the Human Brain*. New York, NJ: G. P. Putnam's Sons.

Damasio, A. R. (1994) *The Feeling of What Happens, Body and Emotion in the Making of Consciouness.* New York, NJ : Harcourt Brace & Company.

Dewey, J. (1932) *Ethics* in *The Collected Works of John Dewey*, (1969-1991) edited by Jo Ann Boydston. Carbondale, IL: Southern Illinois University Press.

Stephen, J. G. (2000) Deconstructing the 'Science Wars' by Reconstructing an Old Mold. *Science,* 287, 253-261.

Investigating Dissociations Between Perceptual Categorization and Explicit Memory

Marci A. Flanery (marci.flanery@vanderbilt.edu)
Thomas J. Palmeri (thomas.j.palmeri@vanderbilt.edu)
Brooke L. Schaper (brooke.l.schaper@vanderbilt.edu)
Department of Psychology; Vanderbilt University
Nashville, TN 37240 USA

Abstract

Are dissociations between categorization and explicit memory in tests of amnesics and normals evidence for multiple memory systems? Or could these dissociations be artifacts arising from methodologies used in some experiments? We report a series of studies exploring this issue. Using normals in various states of simulated amnesia we show that categorization at test is well above chance even in the absence of prior exposure to category members. We also show that subjects perform well when tested with items that conflict with categories they had studied earlier. We argue that subjects in some paradigms can extract information about categories from the test rather than rely on memory for studied category members. In further studies, we generalize these findings to other stimuli and other category structures that have been used in tests of amnesics and normals.

Introduction

Do categorization and explicit memory rely on independent neural memory systems? Evidence for multiple systems comes from dissociations between categorization and explicit memory in studies of normals and amnesics. Amnesics are reported to categorize at levels comparable to normals but are significantly worse at explicit memory. Such dissociations seem to imply that separate systems may exist and pose clear problems for theories that assume a single underlying memory system, such as well-known exemplar models.

The evidence is clear that amnesics have impaired explicit or declarative memory. The focus of this paper is on whether data from studies testing amnesics clearly provide evidence for intact abilities to learn new perceptual categories. Our goal is to examine whether some categorization performance can be explained in the absence of positing a separate implicit system for category learning that is spared in amnesia. Our approach has been to utilize the same paradigms and methodologies found in the amnesia literature to study normal subjects under conditions that simulate aspects of amnesia. To create "amnesia" in normals, we used a variety of techniques such as eliminating the study session altogether, introducing delays between study and test, and surreptitiously switching the test stimuli to those from an unstudied category. We follow the amnesia literature in testing these effects using a variety of stimuli, including distortions of dot patterns, object-like stimuli with discrete features, and simple forms placed in categories separated by quadratic boundaries.

In this paper, we review some behavioral evidence for dissociations from studies of amnesics and normals. For each case, we present data from studies we conducted that provide a possible alternative explanation for intact categorization by amnesics. Due to space constraints, we will only present our results in summarized form without detailed description of the methods or statistical analyses. After summarizing our initial work along these lines reported by Palmeri and Flanery (1999), we describe several new experiments that expanded upon these initial results in several important ways.

Learning Categories of Dot Patterns

A classic methodology for studying categorization and recognition has been the Posner and Keele (1968) dot pattern paradigm. To create a pattern, a small number of dots are randomly scattered on a grid. To create a category, a pattern is randomly generated and designated the prototype. Category members are generated by randomly distorting the prototype by varying degrees.

Knowlton and Squire (1993) used a variant of this paradigm to test amnesics on categorization and recognition. For categorization, subjects were exposed to 40 high distortions. Subjects were tested on judging members and nonmembers of that category. For categorization, members were 4 repetitions of the prototype, 20 low distortions, and 20 high distortions. Nonmembers were 40 randomly generated patterns. For recognition, subjects were exposed to five random patterns eight times each. In the recognition test, they were asked to discriminate between the five old patterns and five new patterns. No corrective feedback was provided in either condition. Knowlton and Squire (1993) reported a dissociation between categorization and recognition when comparing amnesics and normals. As shown in Figure 1, amnesics categorized as well as normals but were significantly impaired at recognition memory.

This dissociation seemed to provide evidence for separate systems. However, Nosofsky and Zaki (1998) showed that a single-system model could account for a

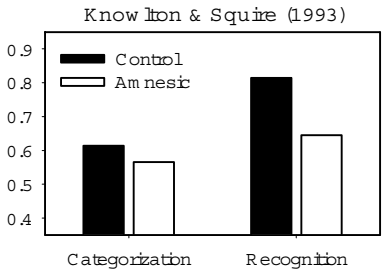


Figure 1. Categorization and recognition accuracy for controls and amnesics.

dissociation by simply assuming that amnesics had degraded memory. A challenge for this theoretical possibility was an extreme dissociation observed by Squire and Knowlton (1995). Their patient, E.P., was able to categorize as well as normals but recognition was entirely at chance. It would be very difficult to formulate a single-system model along the lines of Nosofsky and Zaki that could predict chance recognition performance in the presence of normal categorization performance.

To better understand the categorization performance of amnesics, Palmeri and Flanery (1999) investigated whether prior exposure was even necessary to categorize at test. One explanation for above-chance categorization by amnesics is that it may be possible to group items during the test that looked similar (prototypes and distortions) into the member category and group items that did not look similar (random patterns) into the nonmember category. Whereas, it is impossible to tell apart old from new patterns without memory.

Palmeri and Flanery tested this possibility by producing a state of profound amnesia in normals. As a ruse, subjects were told that patterns had been subliminally presented during an initial word identification task. No dot patterns were ever really presented. Subjects then completed the same categorization and recognition tests used by Knowlton and Squire. Similar to E.P., our simulated profound amnesics showed chance recognition. Yet, they showed above chance categorization. Apparently, our subjects were able to figure out how to categorize members versus nonmembers by picking up on the category structure clearly embedded within the test. They had no prior memories of any sort to rely on.

Experiment 1. We extended this paradigm by directly comparing the performance of simulated amnesics (No Exposure) to that of subjects who were exposed to the study items (Exposure). Half of the subjects were given subliminal exposure, as in Palmeri and Flanery, and were tested on categorization or recognition; the other half were given actual exposure, as in Knowlton and Squire, and tested on categorization or recognition. Results are shown in Figure 2. As expected, the exposure group could recognize items well above chance but the no exposure group could only guess. Replicating Palmeri and Flanery (1999), subjects in a no exposure group could categorize well above chance. Subjects receiving no exposure did not categorize significantly worse than subjects who were actually exposed to category items. Apparently, prior exposure to a category did not provide much, if any, benefit.

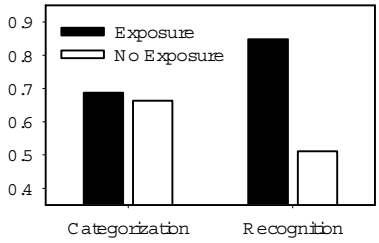


Figure 2. Categorization and recognition accuracy in Experiment 1 for subjects exposed and not exposed to category items (Palmeri & Flanery, 2001).

Experiment 2. One criticism of these studies is that the ruse used to induce amnesia may have placed subjects in a different mind set from that of subjects who were exposed to members. Our "profound amnesics" may realize they never saw any patterns and may think the task is to discover the hidden category structure, something they appear to do quite ably. So, one goal of this experiment was to use a different paradigm for demonstrating that subjects may categorize based on information they acquire during the categorization test. In this experiment, we surreptitiously switched the test stimuli for some subjects to that of an unstudied category.

In addition, we clearly do not want to draw the conclusion that people always ignore information about a previously studied category in favor of information presented during a test. A second goal was to show that when initial exposure provides evidence for a clear category structure, subjects will use that information to make category decisions irrespective of the makeup of the categorization test. To demonstrate this, we adapted a paradigm used by Squire and Knowlton (1995). In one condition, subjects were initially exposed to 40 high distortions of the prototype (40H), exactly as was done in earlier studies. In another condition, subjects were instead exposed to 40 repetitions of the prototype (40P). We reasoned that subjects in the 40P condition should have acquired clear knowledge of the category structure and should protest any surreptitious changes during a test. By contrast, subjects in the 40H condition should have acquired little knowledge of the category structure and should go along with our surreptitious changes.

To verify that different exposure conditions had a significant effect on performance, we tested subjects in the same way as our earlier studies after a one week delay. Overall, 40P subjects achieved 75.3% accuracy and 40H subjects achieved 65.1% accuracy. As expected, categorization accuracy was influenced by the type of information presented during initial category exposure, as was reported by Squire and Knowlton (1995). Overall performance of our 40H subjects was quite comparable to what we and others have observed in this para-

digm; performance of the 40P subjects was significantly better than what we have observed before. So, information presented during initial exposure can have a significant effect on categorization performance.

As a way of simulating amnesia, we tested a subset of subjects after a several weeks delay. But now we tested half on items generated from the prototype used to generate items they had seen before (Same condition) and tested half on items generated from a novel prototype (Different condition). Thus, each subject was in one of four conditions: 40P-Same, 40P-Different, 40H-Same, and 40H-Different. Since all subjects viewed a different randomly generated set of stimuli, we can characterize subjects in the Different condition as receiving a categorization test intended for another individual.

As illustrated in Figure 3, we found that subjects in 40P-Same performed quite well, correctly categorizing 77.4% of the items. However, subjects in 40P-Different were completely at chance categorizing the test items. We suspect that these subjects tried to use the category information they clearly had acquired earlier and could not apply that knowledge when given a test comprised of entirely novel items. For subjects in the 40H conditions, as we predicted, there was no significant difference in performance between subjects who were tested on the same structure they were initially exposed to and subjects who were tested on a completely different structure. Consistent with our previous results, these subjects appear to be making categorization decisions based on what they acquired during the categorization test, not on what they may have acquired during earlier phases of the experiment.

Summary. The dissociation between categorization and recognition reported by Knowlton and Squire (1993) initially appeared to present strong evidence supporting multiple memory systems theory. We reported how the observed dissociations between categorization and recognition using distorted dot patterns may be explained as a result of the particular methodologies used to test these individuals. We showed that very good categorization performance can be achieved in the absence of any prior exposure to the category members. We also showed that very good categorization performance can be achieved when people are tested on items that are

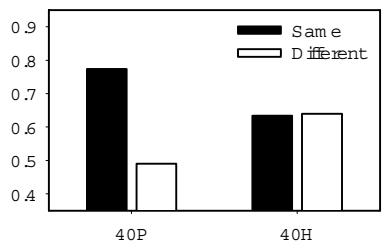


Figure 3. Categorization accuracy in Same condition and Different condition for subjects studying 40 repetitions of a prototype (40P) and 40 high distortions of prototype (40H) in Experiment 2.

different from what they had actually studied. But this seems to only occur when subjects have been initially exposed to a very diffuse category structure consisting of high distortions that are not very similar to one another. When subjects have been exposed to a clear category structure through repetition of a single prototype, they attempt to categorize items based on that acquired category knowledge, not on information presented during the categorization test.

Learning Categories of Object-Like Stimuli with Discrete Features

Reed et al. (1999) aimed to generalize the investigation of preserved categorization by amnesics by using object-like stimuli with discrete features. The stimuli they used, called Peggles, were drawings of animals that varied on nine binary-valued dimensions. To create a category, some Peggle was designated the prototype. Category members were distortions of the prototype. Low distortions shared 7 or 8 features of the prototype. High distortions shared only 1 or 2 features. In the extreme, an antiprototype had all nine features opposite to that of the prototype. Stimuli that shared 4 or 5 features of the prototype were designated neutral items that were half way between the prototype and the antiprototype.

Subjects viewed 40 low distortions of the prototype. After exposure, subjects were told that the animals they saw were all members of a category, called Peggles, and were then asked to make member/nonmember judgments, without feedback, of 96 new items. The test included 12 repetitions of the prototype, 24 low distortions, 24 neutral items, 24 high distortions, and 12 repetitions of the antiprototype. Subjects were also tested on their ability to complete a cued-recall test identifying both values of the 9 dimensions of the Peggles.

Reed et al. (1999) found that amnesics were impaired at an explicit cued-recall task but could categorize at levels comparable to normals. But, two of their amnesics actually categorized stimuli opposite to the way they should have. That is, they mistakenly called the prototype and low distortions nonmembers and called the antiprototype and high distortions members. Reed et al. suggested that amnesics had a spared implicit category learning system that partitioned members and nonmembers but that perhaps declarative memory was needed to remember which partition corresponded to the items they had previously been exposed to.

Experiment 3. Following the theme of this paper, we propose an alternative explanation. During the categorization test, subjects were shown the prototype many times and were shown low distortions that were very similar to the prototype. They were also shown the antiprototype many times and were shown high distortions that were very similar to the antiprototype. In other words, there were two clear clusters of items presented during the categorization test. If subjects could pick up on the category structure embedded within the testing

sequence to cluster stimuli into two groups, they would be able to correctly partition the stimuli into two categories. But, they would not be able to unambiguously decide which cluster corresponded to the category they were initially exposed to without relying on memory of some sort. Might this be a more reasonable explanation of the category switching by amnesics previously reported by Reed et al.?

The goal of this experiment was to test whether subjects might be categorizing in part by extracting information from the structure of the categorization test. We tested subjects in three conditions: Immediate, Delayed, and Novel. The Immediate condition was essentially a replication of Reed et al. (1999). In the Delayed condition, subjects were exposed to the category and then returned one week later to be tested in the same way as subjects in the Immediate condition. In the Novel condition, subjects were also exposed to the category and returned one week later. The stimuli presented for categorization in the Novel condition contained an embedded category structure that contradicted what was presented during initial exposure. To do this, a neutral item with respect to the prototype that was used to generate stimuli from the original exposure session was picked at random and designated the "prototype" for purposes of creating a new categorization test sequence. From this novel prototype, low distortions, neutral items, high distortions, and an antiprototype were created. Note that the "antiprototype" for this new structure would also be considered a neutral item with respect to the prototype that was used to generate items subjects were originally exposed to. The novel categorization test consisted of 12 repetitions of the novel prototype, 24 low distortions, 24 neutral items, 24 high distortions, and 12 repetitions of the novel antiprototype.

Let us generate some predictions for the Novel condition. If subjects are categorizing based on what they had been previously exposed to, they should categorize the "prototype" and the "antiprototype" in this novel test sequence equally, as half way between the member and nonmember category with respect to what they had originally studied. However, if subjects are instead picking up on the clear category structure embedded within this novel test sequence, they should group the "prototype" and its distortions in one category and group the "antiprototype" and its distortions in another category. Half of the subjects would call the "prototype" group the members and half would call the "antiprototype" group the members.

Now to the results. First, as shown in the right portion of Figure 4, performance in the recall task was significantly impaired in the Delayed and Novel condition compared to the Immediate condition. Also, as shown in the left of Figure 4, subjects in the Immediate and Delayed conditions showed comparable categorization.

Scoring categorization data for subjects in the Novel condition was somewhat more complicated (Palmeri & Flanery, 2001). Essentially, what we first did was to

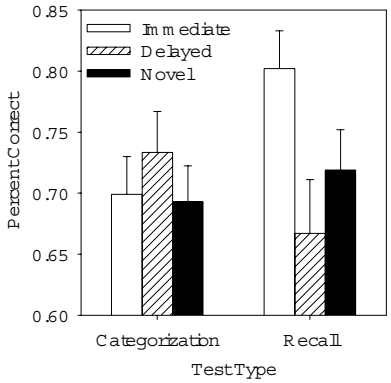


Figure 4. Categorization and cued-recall accuracy in the Immediate, Delayed, and Novel conditions of Experiment 3 (Palmeri & Flanery, 2001).

measure the difference in membership endorsements for the "prototype" and the "antiprototype." Recall that if subjects were categorizing these two critical stimuli with respect to what they had actually been exposed to, they should be indifferent at categorizing these items as members or nonmembers. To the contrary, we found a 53.6% difference in membership endorsements for the "prototypes" and the "antiprototypes." Subjects were clearly discriminating between these items when making category member judgments. Next, if a particular subject judged the "prototype" more often to be a member then we judged categorizations of the low distortions as members and high distortions as nonmembers to be "correct" responses; if a particular subject judged the "antiprototype" more often to be a member then we judged categorizations of the high distortions as members and low distortions as nonmembers to be "correct" responses. Figure 4 displays categorization accuracy for the Novel condition using this scoring method (actually, we scored the Immediate and Delayed conditions in the same way to make the reported results consistent). What should be clear from the figure is that subjects in the Novel condition discriminated between members and nonmembers in a way that was consistent with the structure embedded within the testing sequence and not on memory for what they had seen a week earlier. As with Experiment 2, we found comparable performance between subjects who were tested on categories they actually studied and subjects who were tested on categories that contradicted what they had actually studied.

Summary. In this experiment, we extended a paradigm used by Reed et al. (1999) to contrast categorization and recall by amnesics and normals. They observed impairments in cued recall by amnesics compared to normals, but there was little difference in categorization between the two groups. However, they did observe that two of their amnesic individuals categorized members of the previously studied category as nonmembers and nonmembers as members. While Reed et al. interpreted these results in terms of an implicit memory for the category, we instead provided evidence that this ability

to discriminate members from nonmembers might emerge from a clear category distinction embedded within the testing sequence.

Learning Categories Described by a Complex Quadratic Rule

Can individuals with explicit memory impairments learn to categorize stimuli in accordance with a complex categorization rule? Filoteo et al. (in press) had normals and amnesics learn categories described by what they characterized as a complex quadratic rule. Subjects learned two categories that were defined by multivariate normal distributions. Figure 5 displays the equal likelihood contours for the category structures utilized by Filoteo et al. (in press). Because the categories are defined by normal distributions, the two categories overlap, so perfect performance is impossible. Also as shown in the figure, learning the categories required subjects to integrate information across both stimulus dimensions; in the language of decision boundary theory, learning these categories required the formation of a quadratic (nonlinear) decision rule. This manipulation was of theoretical importance because some work has suggested that amnesics cannot integrate information across multiple dimensions (Rickard & Grafman, 1998).

The stimuli used by Filoteo et al. (in press) consisted of a horizontal and a vertical line connected at the top left corner. The length of the horizontal and vertical lines varied in accordance with the distributions shown in Figure 5. It is important to note that the "diagonal" distribution consisted of stimuli for which the line lengths are highly correlated; in other words, they form the left and top portions of a square (square category). The "circular" distribution consisted of stimuli for which the line lengths are uncorrelated; in other words, they form the left and top portions of various rectangles (rectangle category). On each trial of the experiment, subjects were presented with a stimulus randomly drawn from either the square or the rectangle category, categorized it as a member of category A or category B, and received corrective feedback.

Filoteo et al. observed the accuracy in the last 100 trials to be 85% for normals and 84% for amnesics. They concluded that amnesics appear to be able to acquire categories defined by a complex quadratic rule. To test whether an amnesic could retain that rule over a delay period, they tested one amnesic and one normal after a one day delay. Subjects completed a single block of 100 trials in which they received corrective feedback on every trial, just as in the original training. Accuracy was 92% for the normal individual and 89% for the amnesic. Thus, amnesics and normals appear to be able to learn and retain a quadratic categorization rule.

Experiment 4. The Filoteo et al. results suggest that amnesics can learn and retain a category described by a complex quadratic rule that requires integrating information from two stimulus dimensions, height and width. However, these stimuli can also be described in an alternative way by rotating the dimensions by 45 degrees. That is, we can alternatively describe the dimensions as shape and size. The square and rectangle categories contain stimuli of the same shape and can be categorized by a very simple shape rule rather than a complex quadratic rule. Filoteo et al. rejected this possibility, arguing that their subjects were learning a complex quadratic rule requiring an integration of information along two independent stimulus dimensions. But, we are puzzled by how these subjects were able to learn a complex categorization rule so quickly, reaching asymptotic performance after less than 100 trials, when other categorization experiments examining quadratic boundaries have required many days of training to reach asymptote.

To illustrate that subjects may not be learning a complex quadratic rule, but may instead may be learning a simple shape rule, we replicated and extended the Filoteo et al. study using three conditions. In the first condition, we used the same stimuli and category structures as Filoteo et al. (Square/Rectangle condition). In the second condition, subjects were trained on similar stimuli, but both multivariate category distributions were shifted along dimension 1. In this way, the diagonal category distribution still had height and width correlated, but their values were not equal – in other words, the stimuli were rectangles of the same shape that varied in size (Rectangle/Rectangle condition). In the third condition, we used very different stimulus dimensions of circle size and angle of a diameter line (Circle-Line/Circle-Line condition) that cannot be integrated like the height and width of line segments; these dimensions were roughly equated for discriminability with the height and width dimensions.

Performance in the Squares and Rectangles conditions were comparable (81% and 78% accuracy, respectively). Performance in the Circle-Line condition was far worse (58% accuracy). These results suggest that amnesics may not have been learning a complex quadratic categorization rule at all, but may have instead been learning a very simple shape rule.

Another issue with the Filoteo et al. (in press) results regards the retention of the categorization rule after a

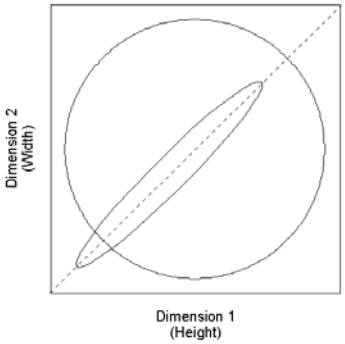


Figure 5. Equal probability contours for categories used by Filoteo et al. and used in Experiment 4.

delay. In the second session of their experiment, subjects received feedback after each trial, similar to what they had experienced during training. Did the amnesics display a real memory for the categorization rule or did they express a savings in relearning a very simple categorization rule? To show that different kinds of categorization tests can reveal different levels of knowledge about categories, we brought our subjects back after one day and tested them in three different ways. First, we tested them without feedback on stimuli drawn from a uniform distribution across the entire set of possible stimuli. Second, we tested them without feedback on stimuli drawn randomly from the two category distributions. Third, we retrained them with feedback, as was done by Filoteo et al. Although subjects reached comparable levels of performance in the Squares and Rectangles condition on the first day, subjects were much better when tested on the uniform distribution without feedback in the Squares condition (96%) than the Rectangles condition (79%). By contrast, in the other two testing conditions (without feedback and with feedback), performance was comparable for the Squares and Rectangles condition (82% and 80% accuracy, respectively). It appears that the different categorization tests can reveal differential knowledge of the categories.

Summary and Conclusions

We found evidence that normal subjects can acquire information about categories in the absence of prior study and in opposition to prior study. In our experiments, performance by subjects in these conditions was not significantly different from performance by subjects who actually received prior study and who were tested on items consistent with their prior study. Our results demonstrate that classification decisions made during a categorization test may not be based solely on information acquired during a study task, but may also be based on information acquired during the test itself. As a general point, we argue that care must be taken in selecting items for a categorization test so as not to provide additional information about the categories being tested or so as not to change the information about the categories that may have been previously acquired. In the present experiments, subjects were tested in such a way that it was possible to extract information about the categories from the tests themselves. Subjects were repeatedly tested on the category prototype (four times in Experiment 1, twelve times in Experiment 2) and were tested on many low distortions that were very similar to the category prototype, conditions particularly amendable to unsupervised category learning. A preferable way to test individuals in an neutral manner might be to sample all possible test stimuli from a uniform distribution, as we did in the last experiment. Although it may be possible to partition such test stimuli into some arbitrary set of categories, only by chance might this partition match the correct category discrimination without relying on memory for studied items.

As we stated at the outset, our results may have implications for understanding the relationship between categorization and other forms of memory. The ability of amnesics to categorize stimuli coupled with their impairment at recognizing or recalling stimuli has been taken as evidence for multiple memory systems (e.g., Knowlton & Squire, 1993; Reed et al., 1999; Squire & Knowlton, 1995; see, however, Nosofsky & Zaki, 1998). If the paradigms used by some investigators permit category acquisition from the categorization test (by contrast, the explicit memory tasks used in these experiments cannot be accurately performed without memory for the studied items), then the strength of this dissociation may be questioned. It seems prudent to forgo strong conclusions about independence or nonindependence of fundamental aspects of human cognition until more convincing paradigms are employed.

Acknowledgments

This work was supported by Vanderbilt University Research Council Grants, Grant BCS-9910756 from the National Science Foundation, and Grant MH61370 from the National Institute of Mental Health.

References

Filoteo, J.V., Maddox, W.T., & Davis, J.D. (in press). Quantitative modeling of category learning in amnesic patients. Journal of the International Neuropsychological Society.

Knowlton, B.J., & Squire, L.R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. Science, 262, 1747-1749.

Nosofsky, R.M., & Zaki, S.R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. Psychological Science, 9, 247-255.

Palmeri, T.J., & Flanery, M.A. (1999). Learning about categories in the absence of training: Profound amnesia and the relationship between perceptual categorization and recognition memory. Psychological Science, 10, 526-530.

Palmeri, T.J., & Flanery, M.A. (2001). Category knowledge acquired during categorization testing.

Posner, M.I., & Keele, S.W. (1968). On the genesis of abstract ideas. Journal of Experimental Psychology, 77, 353-363.

Reed, J.M., Squire, L.R., Patalano, A.L., Smith, E.E., & Jonides, J. (1999) Learning about categories that are defined by object-like stimuli despite impaired declarative memory. Behavioral Neuroscience, 113, 411-419.

Rickard, T.C., & Grafman, J. (1998). Losing their configural mind: Amnesic patients fail on transverse patterning. Journal of Cognition Neuroscience, 10, 509-524.

Squire, L.R., & Knowlton, B.J. (1995). Learning about categories in the absence of memory. PNAS, 92, 12470-12474.

Development of Physics Text Corpora for Latent Semantic Analysis

Donald R. Franceschetti (dfrncsch@memphis.edu)
Department of Physics, University of Memphis, CAMPUS BOX 523390
Memphis, TN 38152 USA

Ashish Karnavat (akarnavat@hotmail.com)
Department of Computer Science, University of Memphis, CAMPUS BOX 526429
Memphis, TN 38152 USA

Johanna Marineau (jmarinea@memphis.edu)
Department of Psychology, 202 Psychology Building
University of Memphis, Memphis, TN 38152 USA

Genna L. McCallie (jordy911@bellsouth.net)
Department of Psychology, 202 Psychology Building
University of Memphis, Memphis, TN 38152 USA

Brent A. Olde (baolde@memphis.edu)
Department of Psychology, 202 Psychology Building
University of Memphis, Memphis, TN 38152 USA

Blair L. Terry (bterry@memphis.edu)
Department of Psychology, 202 Psychology Building
University of Memphis, Memphis, TN 38152 USA

Arthur C. Graesser (a-graesser@memphis.edu)
Department of Psychology, 202 Psychology Building
University of Memphis, Memphis, TN 38152 USA

Abstract

Student responses to qualitative physics questions were analyzed with latent semantic analysis (LSA), using different text corpora. Physics potentially has a number of distinctive characteristics that are not encountered in many other knowledge domains. Physics texts exist at a variety of levels and typically involve an integrated presentation of text, figures and equations. We explore the adequacy of several text corpora and report results on vector lengths and correlations between key terms in elementary mechanics. The results suggest that a carefully constructed smaller corpus may provide a more accurate representation of fundamental physical concepts than a much larger one.

Introduction

The physics classroom has often served as a laboratory for cognitive science. Studies of students learning or failing to learn physics have influenced notions of conceptual change, question answering and tutoring strategy (Albacete & VanLehn, 2000; Van Heuvelen, 1991). The physics teaching community is now aware that conventional teaching methods often fail to make any significant change in the student's understanding of the physical world. While students in the more technical introductory courses might develop the ability to recognize certain problem templates and to manipulate equations, and those in "conceptual" physics courses learn enough set answers to pass multiple choice exams, there is ample evidence that many students retain the same misconceptions about the nature of everyday phenomena with which they began the formal study of physics (Ploetzner & VanLehn, 1997).

The study of physics allegedly places rather different demands on the student than other academic work, as is readily apparent in the texts that are used. The goal of "understanding physics" or "thinking like a physicist", to which most instructors aspire, involves a combination of declarative and procedural knowledge in which the procedural component figures far more significantly than in, for example, a survey of history or introductory computer literacy. Language is used somewhat differently in physics than in other scientific fields. While biology and chemistry resort to Greco-Roman or Germanic word forming conventions to

introduce new words with precise meanings, physics more often than not takes words from ordinary language, like force and momentum, and restricts their meaning to a single sense. In most modern physics texts (such as Hewitt, 1998), there are multiple photographs or simple sketches on every page, and much of the text is directly organized around these figures. Much of the exposition in conceptual physics courses includes questions and answers that may be separated by text. Physics texts often devote considerable space to the historical evolution of physical concepts, the cultural context of physics, and its social impact. Some authors also devote appreciable space to discussing discarded theories and chains of reasoning that lead to incorrect conclusions. Thus, a significant fraction of the text found in a physics text may, in fact, exemplify incorrect thinking.

Our group has been developing a corpus of texts about physics that will eventually be used in an intelligent tutoring system on conceptual physics. The text corpus is needed to build a latent semantic analysis (LSA) space, which will be used to process the meaning of student answers in ordinary language. This paper is concerned particularly with the best strategy to construct such a corpus. A naive approach would be to gather a number of physics texts, and combine them into one corpus. However, there are unusual challenges taking this approach. What should be done about the diagrams in the text? What about the text that was written to illustrate incorrect reasoning? Does the inclusion of texts at different levels strengthen or dilute the accuracy with which physics concepts are represented in the LSA space? In short, how much special preparation of the corpus is needed, if it is to provide a reliable representation of the physics that students are expected to learn?

Latent Semantic Analysis

LSA has recently been proposed as a statistical representation of a large body of world knowledge (Kintsch, 1998; Landauer & Dumais, 1997). LSA provides the foundation for grading essays, even essays that are not well formed grammatically, semantically, and rhetorically; in fact, LSA-based essay graders assign grades to assays as reliably as experts in composition (Foltz, Gilliam, & Kendall, 2000). LSA has been used to evaluate the quality of student contributions in interactive dialogs between college students and AutoTutor, a tutoring system in the domain of computer literacy; the LSA module evaluates the quality of student answers to questions almost as reliably as graduate student research assistants (Graesser, P. Wiemer-Hastings, K. Wiemer-Hastings, Harter, Person, & TRG, 2000; P. Wiemer-Hastings, K. Wiemer-Hastings, Graesser, & TRG, 1999). Given these successes in using LSA to evaluate the quality of student essays and contributions in tutoring systems, on a variety of topics, we were interested in exploring how LSA would fare in the domain of qualitative physics.

LSA is a mathematical technique in which the information contained in the co-occurrences of words in a body of text is compressed into a set of vectors in N-dimensional space. The input to LSA is word co-occurrence matrix M, where the individual elements Mij is the number of times that the ith word occurs in the jth document. A document is an arbitrarily defined unit, but normally is a sentence, paragraph, or section in a text. The rows and columns of the matrix are then subjected to mathematical transformations that take into account the frequency of word use in the document (Berry, Dumais, & O'Brien, 1995; Landauer, Foltz, & Laham, 1998). Using the mathematical process of singular value decomposition, the matrix is then expressed as the product of three matrices, the second of which contains the singular values on the diagonal. Changing all but the largest N singular values to zero sets the dimensionality N of the vector space representing the text. The matrices are then re-multiplied to produce a matrix of the same dimensions of the original matrix.

At first glance it might seem that by discarding some of the singular values we are discarding information. However, it turns out in practice that the lower dimensional representation better captures the meaning of the text. For instance, there ends up being a positive relationship between the coefficients in the rows corresponding to different words, if the words have similar or associated meanings. The reduced number of dimensions are sufficient for evaluating the conceptual relatedness between any two bags of words. A bag is an unordered set of one or more words. The match (i.e., similarity in meaning, conceptual relatedness) between two bags of words is computed as the geometric cosine (or dot product) between the two associated vectors, with values that normally range from 0 to 1. LSA successfully predicts the coherence of successive sentences in text (Foltz, Kintsch, & Landauer, 1998), the similarity between student answers and ideal answers to questions (Graesser, P. Wiemer-Hastings, et al., 2000; Wiemer-Hastings et al., 1999), and the structural distance between nodes in conceptual graph structures (Graesser, Karnavat, Pomeroy, Wiemer-Hastings, & TRG, 2000). At this point, researchers are exploring the strengths and limitations of LSA in representing world knowledge.

Constructing an LSA Corpus That Knows About Physics

We have assembled several different physics corpora to test the effect of the content of the subject matter on the

quality of the LSA solutions. The documents in the texts were classified into different rhetorical categories, such as exposition, example problems, historico-cultural material, incorrect reasoning, and so on. The fundamental research question is whether the inclusion of different texts and categories of content have an impact on the representation of core concepts in the mechanics portion of a conceptual physics course. All the corpora include text materials from the mechanics portion of Paul Hewitt's *Conceptual Physics* (1998), a text that is widely used in conceptual physics courses at the college level; these were used with permission from the publisher. The "Omnibus" corpus included chapters 2-9 of the Hewitt book plus six volumes of a comprehensive text aimed at students in technical or life science majors, two advanced texts in electromagnetism, and another physics text that was available electronically. The "Large" corpus was constructed from the former by deleting the three latter texts. A "Small" corpus further deleted the texts that did not cover mechanics. A "Restricted Small" corpus further deleted any text identified as primarily historico-cultural or involving misconceptions. In the "Restricted Hewitt" corpus, we included only those texts from Hewitt in the restricted small corpus. Each of the corpora was thus a proper subset of the preceding one, with the goal of further refining or sanitizing the text corpus to handle the core concepts in mechanics. The time needed to "restrict" a text was minimal once the text was converted to electronic form.

Vector Lengths and Similarity

Kintsch (1998) proposed that the length of the vectors representing key terms provides a measure of the extent to which the LSA has captured the meaning (or importance, centrality) of the word with respect to the subject matter. The vector length increases to the extent that the set of values in the vector deviate from zero. Words like *force*, *momentum* and *gravitation* should have reasonably large vector lengths in any corpus that represented basic physics concepts well.

LSA spaces of 100, 200, 300, 400, and 500 dimensions were created for each of the above five corpora. Each text paragraph was treated as a document. Figure captions were eliminated. Questions & answers were lumped in the same document. Based on the vector lengths computed for the key mechanics words listed in Table 1, it was decided that little improvement would be achieved by going beyond 500 dimensions. The restricted Hewitt corpus was so small that only a 400 dimensional representation could be obtained. The vector lengths for selected physics words in a 300 dimensional space are shown in Table 1.

A number of conclusions can readily be drawn from Table 1. There is a general correlation between vector lengths of the first two corpora and between those of the two smallest corpora. When we eliminated the material not pertinent to mechanics as presently understood, some vectors ended up increasing in length. For example, the *impulse* concept, which occurs only in mechanics, had a significantly larger vector length in the smaller corpora than in the larger corpora. The same can be said for *tension*, which is the force transmitted by a rope or cable, and is useful only in mechanics. Even a concept like *energy*, which pervades all areas of physics, appears to be more crisply represented in the smaller corpus.

Table 1. Vector Lengths for Physics Words.

Word	Omnibus	Large	Small	R-small	R-Hewitt
Gravity	.288	.281	.262	.242	.240
Gravitational	.256	.250	.223	.256	.283
Mass	.300	.293	.269	.239	.288
Acceleration	.300	.296	.266	.270	.284
Force	.186	.179	.155	.128	.153
Momentum	.267	.263	.258	.283	.288
Energy	.222	.219	.228	.238	.313
Impulse	.367	.371	.400	.432	.466
Friction	.320	.314	.301	.313	.375
Velocity	.252	.250	.240	.236	.291
Vector	.285	.305	.292	.382	.455
Potential	.323	.328	.361	.386	.464
Tension	.266	.271	.302	.390	.475
Kinetic	.298	.294	.301	.312	.422
Normal	.315	.314	.352	.414	.373
Newton	.347	.242	.211	.206	.265
Aristotle	.309	.309	.318	.409	.436
Galileo	.324	.326	.325	.338	.355
Newtonian	.242	.223	.221	.339	.000

The names of the key physicists, *Galileo* and *Newton*, along with that of *Aristotle*, whose notions of physics are now largely discarded, were also included in our study of vector lengths. Interestingly, our efforts to eliminate material of only historical value in the two restricted corpora did not eliminate a rather well represented *Aristotle*. LSA did pick up stylistic characteristics of individual authors.

The similarity between concepts in LSA is represented by the cosine values in corresponding vectors. We computed the cosines between the physics terms in Table 1 and these appear in Table 2. The greatest similarity appeared for *kinetic energy*, which is in effect a composite word and for *impulse-momentum*, which would appear as a composite in the "impulse-momentum theorem" and the exposition of it, in that impulse equals the net change of momentum in a collision. We note that the similarities between *mass* and *acceleration* and between *force* and *acceleration*, which would be expected in any exposition based on Newton's second law (the net force on an object equals

the mass times its acceleration). The similarity scores are appreciably more apparent in the smaller corpora with the irrelevant text removed.

Table 2. Largest magnitude cosines between key physics terms. (Corpora titles are abbreviated)

Correlation	Om	Lge	Sm	R-Sm	R-H
Gravitational force	.084	.083	.093	.146	.029
Gravitational potential	.058	.091	.097	.107	.032
Force acceleration	.006	.009	.009	.048	.087
Mass acceleration	.033	.035	.044	.070	.066
Normal force	.080	.084	.125	.096	.043
Mass momentum	.010	.013	.028	.040	.077
Impulse momentum	.182	.187	.176	.196	.148
Kinetic energy	.209	.228	.265	.267	.267
Tension friction	.052	.052	.020	.066	.001
Vector Velocity	.052	.055	.053	.065	.059
Kinetic friction	.081	.083	.020	.066	.026

Summary

We have developed a number of alternative physics text corpora for use in the evaluation of student answers to physics questions. Comparisons of word length and word similarity suggest that both the elimination of material from other areas of physics and other levels of exposition, as well as the elimination of material not dealing with the exposition of the physical concepts, allows an improved representation of core physics terms and the relationships between them, even with a rather small corpus. However, this conclusion is currently being tested on a large body of student and expert answers to physics questions. The preliminary results suggest that although vector lengths increase for individual words with a refined selection of texts, it is a large corpus that works best when the entire sentence is used to evaluate the match of student and expert answers. In other words, individual words may have a crisper representation when a smaller, well-defined text is used but when analyzing an answer formed around the integration of several complex concepts, a broader selection of texts is more beneficial. Furthermore, it is our contention that a regression could be used to capitalize on the unique information provided by both types of LSA spaces. In future work, we will examine the feasibility of adding picture descriptions in natural language to the corpus and alternative treatments of equations and composite words.

Acknowledgments

This research was supported by Grant N00014-00-1-0600 from the Cognitive Science Division of the Office of Naval Research.

References

Albacete, P. L., & VanLehn, K. A. (2000), Evaluating the effectiveness of a cognitive tutor for fundamental physics concepts. *Proceedings of the22nd Annual Conference of the Cognitive Science Society* (pp. 25-30). Mahwah, NJ: Lawrence Erlbaum.

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995), Using linear algebra for intelligent information retrieval. *SIAM Review, 37*, 573-595.

Foltz, W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments, 8*, 111-128.

Foltz, P. W., Kintsch, W., & Landauer, T. K, (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes 25*, 285-307.

Graesser, A. C., Karnavat, A., Pomeroy, A., Wiemer-Hastings, K., & TRG (2000), Latent semantic analysis captures causal, goal-oriented, and taxonomic structures. *Proceedings of the22nd Annual Conference of the Cognitive Science Society* (pp. 184-189) Mahwah, NJ: Erlbaum.

Graesser, A.C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., & Person, N., and the TRG (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments, 8*, 128-148.

Hewitt, P. G. (1998) *Conceptual physics* (Ed. 8). Reading, MA: Addison Wesley Longman.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, MA: Cambridge University Press.

Landauer, T. K., & Dumais, S. T. (1997) A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284.

Ploetzner, R., & VanLehn, K. (1997). *Cognition & Instruction, 15*, 169-205.

Van Heuvelen, A. (1991). Learning to think like a physicist: A review or research-based instructional strategies, *American Journal of Physics, 59*, 891-897.

Wiemer-Hastings, P., Wiemer-Hastings, K., Graesser, A. & TRG (1999). Improving an intelligent tutor's comprehension of students with latent semantic analysis. In S. Lajoie & M. Vivet (Eds.), *Artificial intelligence in education* (pp. 535-542). Amsterdam: IOS Press.

Modeling Cognition with Software Agents

Stan Franklin (franklin@memphis.edu)
Institute for Intelligent Systems, The University of Memphis, Memphis TN 38152, US

Art Graesser (a-graesser@memphis.edu)
Institute for Intelligent Systems, The University of Memphis, Memphis TN 38152, US

Abstract

We propose the use of autonomous software agents as cognitive models that generate testable hypotheses about human cognition. While such agents are typically produced to automate practical human tasks, they can be designed within the constraints of a psychological theory. As an example we describe an agent designed within global workspace theory that accommodates several other theories as well. We discuss various resulting hypotheses, including a new interpretation of the readiness potential data of Libet.

Introduction

Computational models have long been a major, and perhaps indispensable, tool in cognitive science. Many of these model some psychological theory of a particular aspect of cognition, attempting to account for experimental data. Others aspire to be a general computational model of cognition, such as the construction-integration model (Kintsch 1998), SOAR (Laird et al. 1987), and ACT-R (Anderson 1990). Most of these computational models are computer simulations of subjects in psychological laboratories, and are capable of performing tasks at a fine-grain level of detail. The simulated data ideally fit the human data like a glove. The theories on which the simulations are based are periodically revised so that new simulations conform more closely to the data. The computational models are judged on how closely they predict the data. A model may also be judged by the amount of change required in core, as opposed to peripheral, parameters that are needed to fit the data. Alternatively, the models are evaluated on a course-grain level, by observing whether a number of qualitative predictions (i.e., directional predications, such as condition A > B) fit the data. And finally, all of the models have been evaluated by observing how well they fit data in practical, everyday tasks in real-world environments. For example, some such agents, based on SOAR, simulate battlefield performers such as fighter pilots and tank commanders (Hirst & Kalus 1998). These data fitting approaches to testing theories have been hugely successful, and account for a large body of what is now known in cognitive science.

In this paper, we propose another class of computational models, which fall under the general heading of autonomous software agents (Franklin & Graesser 1997). These agents are designed to implement a theory of cognition and attempt to automate practical tasks typically performed by humans. We have been developing two such agents that implement global workspace theory Baars 1988), one with a relatively simple clerical task (Zhang et al. 1998b) and the other with a rather complex personnel assignment task (Franklin et al. 1998). These models do not merely produce output that solves a specific engineering problem, as do typical software agents like web bots. They have mechanisms that simulate human cognition and their design decisions generate hopefully testable hypotheses (Franklin 1997), thus potentially providing research direction for cognitive scientists and neuroscientists.

This paper briefly describes the architecture and mechanisms of one such agent. In Table 1 we point out examples of relevant hypotheses that arise from our design decisions. It is beyond the scope of this article to specify all of the hypotheses and associated support.

Theoretical Frameworks

According to *global workspace (GW) theory* (Baars 1988), one principal function of consciousness is to recruit the relevant resources needed for dealing with novel or problematic situations. These resources may include both knowledge and procedures. They are recruited internally, but partially driven by stimulus input. GW theory postulates that human cognition is implemented by a multitude of relatively small, special purpose processes, almost always unconscious. Communication between them is rare since they mostly communicate through working memory and over a narrow bandwidth. They are individually quite simple and incapable of dealing with complex messages. Coalitions of such processes compete for access to a global workspace. This limited capacity workspace serves to broadcast the message of the coalition to all the unconscious processors (bringing it to consciousness) in order to recruit relevant processors to join in handling the current novel situation, or in solving the current problem. Thus consciousness allows us to deal with novel or problematic situations that cannot be dealt with efficiently, if at all, by automated unconscious processes. Consciousness recruits appropriately useful resources, and thereby manages to solve the relevance problem.

An *autonomous agent* (Franklin & Graesser 1997) is a system situated in, and part of, an environment, which senses that environment, and acts on it, over time, in pursuit of its own agenda. In biological agents, this agenda arises from drives that evolve over generations; in artificial agents its designer builds in the drives. Such drives, which act as motive generators (Sloman 1987), must be present, whether explicitly represented or derived from the processing trajectory. The agent also acts in such a way as to possibly influence what it senses at a later time. In other words, it is structurally coupled to its environment (Maturana 1975). Examples include humans, most animals some mobile robots, and various computational agents, including artificial life agents, software agents and many computer viruses. Here we are immediately concerned with autonomous software agents, designed for specific tasks, and 'living' in real world computing systems such as operating systems, databases, or networks.

A *"conscious" software agent* is one that implements GW theory. In addition to modeling this theory (Franklin & Graesser 1999), such "conscious" software agents should be capable of more adaptive, more human-like operations, including being capable of creative problem solving in the face of novel and unexpected situations. However, there is no claim that the agent is a sentient being. What, if anything, the agent truly feels or what the conscious experience actually is are not the relevant concerns.

IDA (Intelligent Distribution Agent) is a "conscious" software agent being developed for the US Navy (Franklin et al. 1998). At the end of each sailor's tour of duty, the sailor is assigned to a new billet.. The Navy employs some 280 people, called detailers, to effect these new assignments. IDA's task is to completely automate the role of detailer. IDA must communicate with sailors via email in natural language, by understanding the content and producing life-like responses. Sometimes she will initiate conversations. She must access several databases, again understanding the content. She must adher to some ninety Navy policies. She must hold down moving costs, but also cater to the needs and desires of the sailor. This includes negotiating with the sailor and eventually writing the orders. A partial prototype of IDA with most of the functionality described is now up and running. It should be complete before the beginning of the year.

Architecture and Mechanisms

Table 1 specifies several of the underlying hypotheses that guided the design of IDA Many of these hypotheses are not directly addressed in this paper. Others will be discussed in some detail.

IDA is intended to model a broad range of human cognitive function. Her architecture is comprised of modules each devoted to a particular cognitive process. Table 2 lists most of these modules and gives pointers to the sources of their computational mechanisms, and to the psychological theories they support.

The processors postulated by GW theory are implemented by codelets, small pieces of code, each an independent thread. These are specialized for some simple task and often play the role of demons waiting for appropriate conditions under which to act. From a biological point of view, these codelets may well correspond to Edelman's neuronal groups (1987).

Perception in IDA consists mostly of processing incoming email messages in natural language. In sufficiently narrow domains, natural language understanding may be achieved via an analysis of surface features called complex, template-based matching (Allen 1995, Jurafsky & Martin 2000). Ida's relatively limited domain requires her to deal with only a few dozen or so distinct message types, each with relatively predictable content. This allows for surface level natural language processing. Her language-processing module has been implemented as a Copycat-like architecture (Hofstadter & Mitchell 1994) with codelets that are triggered by surface features. The mechanism includes a slipnet that stores domain knowledge, a pool of codelets (processors) specialized for recognizing particular pieces of text, and production templates for building and verifying understanding. Together they allow her to recognize, categorize and understand. IDA must also perceive content read from databases, a much easier task. An underlying hypothesis motivating our design decisions about perception appears in Table 1.

Suppose, for example, that IDA receives a message from a sailor saying that his projected rotation date (PRD) is approaching and asking that a job be found for him. The perception module would recognize the sailor's name and social security number, and that the message is of the please-find-job type. This information would then be written to working memory. The hypothesis here is that the contents of perception are written to working memory before becoming conscious. IDA employs sparse distributed memory (SDM) as her major associative memory (Kanerva 1988). SDM is a content addressable memory that, in many ways, is an ideal computational mechanism for use as a long-term associative memory (LTM). Any item written to working memory cues a retrieval from LTM, returning prior activity associated with the current entry. In our example, LTM will be accessed as soon as the message information reaches the workspace, and the retrieved associations will be also written to the workspace.

Table 1. Hypotheses from Design Decisions

Module	Hypotheses from Design Decisions
Perception	Much of human language understanding employs a combined bottom up/top down passing of activation through a hierarchical conceptual net, with the most abstract concepts in the middle.
Working Memory	The contents of perception are written to working memory before becoming conscious.
Long-term Memory	Part, but not all, of working memory, the focus, is set aside as an interface with long-term associative memory (LTM). Reads from LTM are made with cues taken from the focus and the resulting associations are written there. Writes to LTM are also made from the focus.
Consciousness	Human consciousness must have a mechanism for gathering processors (neuronal groups) into coalitions, another for conducting the competition, and yet another for broadcasting
Motivation	The hierarchy of goal contexts is fueled at the top by drives, that is by primitive motivators, and at the bottom by input from the environment, both external and internal
Goal Contexts	In humans, processors (neuronal groups) bring perceptions and thoughts to consciousness. Other processors, aware of the contents of consciousness, instantiate an appropriate goal context hierarchy, which motivates yet other processors to perform internal or external actions.
Emotions	Action selection will be influenced by emotions via their effect on drives. Emotions also influence attention and the strength with which items are stored in associative memory.
Voluntary Action	Voluntary action in humans is controlled by a timekeeper who becomes less patient as the time for a decision increases. Each time a proposal or objection reaches consciousness, its chance of becoming conscious again diminishes.
Language Production	Much of human language production results from filling in blanks in scripts, and concatenating the results.

At a given moment IDA's workspace may contain, ready for use, a current entry from perception or elsewhere, prior entries in various states of decay, and associations instigated by the current entry, i.e. activated elements of LTM.. IDA's workspace thus consists of both short-term working memory (STM) and something very similar to the long-term working memory (LT-WM) of Ericsson and Kintsch (1995).

Since most of IDA's cognition deals with performing routine tasks with novel content, most of her workspace is structured into registers for particular kinds of data. Part of the workspace, the *focus*, is set aside as an interface with long-term LTM. Retrievals from LTM are made with cues taken from the focus and the resulting associations are written to other registers in the focus. The contents of still other registers in the focus are stored in (written to) associative memory. All this leads to the perception hypothesis in Table 1.

Not all of the contents of the workspace eventually make their way into consciousness. The apparatus for "consciousness" consists of a coalition manager, a spotlight controller, a broadcast manager, and a collection of attention codelets who recognize novel or problematic situations (Bogner et al. 2000).

Each attention codelet keeps a watchful eye out for some particular situation to occur that might call for "conscious" intervention. In most cases the attention codelet is watching the workspace, which will likely contain both perceptual information and data created internally, the products of "thoughts." Upon encountering such a situation, the appropriate attention codelet will be associated with the small number of codelets that carry the information describing the situation. (In the example of our message, these codelets would carry the sailor's name, his or her social security number, and the message type.) This association should lead to these information codelets, together with the attention codelet that collected them, becoming a coalition. Codelets also have activations measuring their current relevance. The attention codelet increases its activation in order that the coalition might compete for the spotlight of "consciousness". Upon winning the competition, the contents of the coalition is then broadcast to all codelets. This leads us to the consciousness hypothesis in Table 1.

Baars addresses the question of how content arrives in consciousness (1988, pp. 98-99), offering two possible high-level mechanisms both consistent with neurophysiological timing findings. He also devotes an entire chapter (1988 Chapter 3) to neurophysiological evidence consistent with the basic concept of a global workspace. Yet no mechanisms are proposed for the three distinct processes identified as being needed in our hypothesis above. Here we have a good example of engineering, as well as psychological, considerations giving direction to neurophysiological research.

Summarizing our example, an attention codelet will note the please-find-job message type, gather information codelets carrying name, ssn and message type, be formed into a coalition, and will compete for consciousness. If or when successful, its contents will be broadcast.

IDA depends on a behavior net (Maes 1989) for high-level action selection in the service of built-in

drives. She has several distinct drives operating in parallel that vary in urgency as time passes and the environment changes. Behaviors are typically mid-level actions, many depending on several behavior codelets for their execution. A behavior net is composed of behaviors, corresponding to goal contexts in GW theory, and their various links. A behavior looks very much like a production rule,

Table 2. IDA's Modules and Mechanisms and the Theories they Accommodate

Module	Computational Mechanism motivated by	Theories Accommodated
Perception	Copycat architecture (Hofstadter & Mitchell 1994)	Perceptual Symbol System (Barsalou 1999)
Working Memory	Sparse Distributed Memory (Kanerva 1988)	Long-term Working Memory (Erricsson & Kintsch 1995)
Emotions	Neural Networks (McCellland & Rumelhart 1986)	(Damasio 1999, Rolls 1999)
Associative Memory	Sparse Distributed Memory (Kanerva 1988)	
Consciousness	Pandemonium Theory (Jackson 1987)	Global Workspace Theory (Baars 1988)
Action Selection	Behavior Nets (Maes 1989)	Global Workspace Theory (Baars 1988)
Constraint Satisfaction	Linear Functional (standard operations research)	
Deliberation	Pandemonium Theory (Jackson 1987)	Human-Like Agent Architecture (Sloman 1999)
Voluntary Action	Pandemonium Theory (Jackson 1987)	Ideomotor Theory (James 1890)
Language Generation	Pandemonium Theory (Jackson 1987)	
Metacognition	Fuzzy Classifers	Human-Like Agent Architecture (Sloman 1999)

having preconditions as well as additions and deletions. It typically requires the efforts of several codelets to effect its action.

. Each behavior occupies a node in a digraph. As in connectionist models (McClelland et al. 1986), this digraph spreads activation. The activation comes from that stored in the behaviors themselves, from the environment, from drives, and from internal states. More relevant behaviors receive more activation from the environment. Each drive awards activation to those behaviors that will satisfy it. Certain internal states of the agent can also activate behaviors. One example might be activation from a coalition of codelets responding to a "conscious" broadcast. Activation spreads from behavior to behavior along both excitatory and inhibitory links and a behavior is chosen to execute based on activation. IDA's behavior net produces flexible, tunable action selection. This hierarchy of goal contexts is fueled at the top by drives, that is, by primitive motivators, and at the bottom by input from the environment, both external and internal.

Returning to our example, the broadcast is received by appropriate behavior-priming codelets who know to instantiate a behavior stream for reading the sailor's personnel record. They also bind appropriate variables with name and ssn, and send activation to a behavior that knows how to access the database. When that behavior is executed, behavior codelets associated with it begin to read data from the sailor's file into. the workspace. Each such write results in another round of associations, the triggering of an attention codelet, the resulting information coming to "consciousness," additional binding of variables and passing of activation, and the execution of the next behavior. As long as it's the most important activity going, this process is continued until all the relevant personnel data is written to the workspace. In a similar fashion, repeated runs through "consciousness" and the behavior net result in a course selection of possible suitable jobs being made from the job requisition database.

The process just described leads us to speculate that in humans, like in IDA, processors (neuronal groups) bring perceptions and thoughts to consciousness. Other processors, aware of the contents of consciousness, instantiate an appropriate goal context hierarchy, which in turn, motivates yet other processors to perform internal or external actions.

IDA is provided with a constraint satisfaction module designed around a linear functional. It provides a numerical measure of the suitability, or fitness, of a specific job for a given sailor. For each issue (say moving costs) or policy (say sea duty following shore duty) there's a function that measures suitability in that respect. Coefficients indicate the relative importance of each issue or policy. The weighted sum measures the job's fitness for this sailor at this time. The same process, beginning with an attention codelet and ending with behavior codelets, brings each function value to "consciousness" and writes the next into the workspace. At last, the job's fitness value is written to the workspace.

Since IDA's domain is fairly complex, she requires *deliberation* in the sense of creating possible scenarios, partial plans of actions, and choosing between them (Sloman 1999). In our example, IDA now has a list of a number of possible jobs in her workspace, together with their fitness values. She must construct a temporal scenario for at least a few of these possible billets to see if the timing will work out (say if the sailor can be

aboard ship before the departure date). In each scenario the sailor leaves his or her current post during a prescribed time interval, spends a specified length of time on leave, possibly reports to a training facility on a certain date, uses travel time, and arrives at the new billet with in a specified time frame. Such scenarios are valued on how well they fit the temporal constraints (the gap) and on moving and training costs. These scenarios are composed of scenes organized around events, and are constructed in the workspace by the same process of attention codelet to "consciousness" to behavior net to behavior codelets as described previously.

We humans most often select actions subconsciously, but we also make voluntary choices of action, often as a result of the kind of deliberation described above. Baars argues that such voluntary choice is the same a conscious choice (1997, p. 131). We must carefully distinguish between being conscious of the results of an action and consciously deciding to take that action, that is, of consciously deliberating on the decision. The latter case constitutes voluntary action. William James proposed the *ideomotor theory* of voluntary action (James 1890). James suggests that any idea (internal proposal) for an action that comes to mind (to consciousness) is acted upon unless it provokes some opposing idea or some counter proposal. GW theory adopts James' ideomotor theory as is (1988, Chapter 7), and provides a functional architecture for it. The IDA model furnishes an underlying mechanism that implements that theory of volition and its architecture in a software agent.

Suppose that in our example at least one scenario has been successfully constructed in the workspace. The players in this decision making process include several proposing attention codelets and a timekeeper codelet. A proposing attention codelet's task is to propose that a certain job be offered to the sailor. Choosing a job to propose on the basis of the codelet's particular pattern of preferences, it brings information about itself and the proposed job to "consciousness" so that the timekeeper codelet can know of it. Its preference pattern may include several different issues (say priority, moving cost, gap, etc) with differing weights assigned to each. For example, our proposing attention codelet may place great weight on low moving cost, some weight on fitness value, and little weight on the others. This codelet may propose the second job on the scenario list because of its low cost and high fitness, in spite of low priority and a sizable gap. If no other proposing attention codelet objects (by bringing itself to "consciousness" with an objecting message) and no other such codelet proposes a different job within a prescribed span of time, the timekeeper codelet will mark the proposed job as being one to be offered. If an objection or a new proposal is made in a timely fashion, it will not do so.

Two proposing attention codelets may alternatively propose the same two jobs several times. Several mechanisms tend to prevent continuing oscillation. Each time a codelet proposes the same job it does so with less activation and, so, has less chance of coming to "consciousness." Also, the timekeeper loses patience as the process continues, thereby diminishing the time span required for a decision. A job proposal may also alternate with an objection, rather than with another proposal, with the same kinds of consequences. These occurrences may also be interspersed with the creation of new scenarios. If a job is proposed but objected to, and no other is proposed, the scenario building may be expected to continue yielding the possibility of finding a job that can be agreed upon.

Experimental work of neuroscientist Benjamin Libet lends support to this implementation of voluntary action as mirroring what happens in humans (Libet et al. 1983). He writes, "Freely voluntary acts are preceded by a specific electrical change in the brain (the 'readiness potential', RP) that begins 550 ms before the act. Human subjects became aware of intention to act 350-400 ms after RP starts, but 200 ms. before the motor act. The volitional process is therefore initiated unconsciously. But the conscious function could still control the outcome; it can veto the act." Libet interprets the onset of the readiness potential as the time of the decision to act. Suppose we interpret it, instead, as the time a neuronal group (attention codelet) decides to propose the action (job). The next 350-400 ms would be the time required for the neuronal group (attention codelet) to gather its information (information codelets) and win the competition for consciousness. The next 200 ms would be the time during which another neuronal group (timekeeper) would wait for objections or alternative proposals from some third neuronal group (attention codelet) before initiating the action. This scenario gets the sequence right, but begs the question of the timing. Why should it take 350 ms for the first neuronal group (attention codelet) to reach consciousness and only 200 ms for the next? Our model would require such extra time during the first pass to set up the appropriate goal context hierarchy (behavior stream) for the voluntary decision making process, but would not require it again during the second. The problem with this explanation is that we identify the moment of "consciousness" with the broadcast, which occurs before instantiation of the behavior stream. So the relevant question is whether consciousness occurs in humans only after a responding goal structure is in place? This leads us to the voluntary action hypothesis in Table 1.

Future Work

Though the IDA model cuts a broad swath, human cognition is far too rich to be easily encompassed. Still, we plan to extend the model in several ways. An alteration to the behavior net will allow automation of actions. A capacity for learning from conversations with detailers is planned (Ramamurthy et al. 1998). A development/training period utilizing that ability is also anticipated for IDA (Franklin 2000). We're also working on giving her the ability to report "conscious" activity in natural language. Though IDA deals intelligently with novel instances of routine situations, she should be able to also handle unexpected, and problematic non-routine situations. We're working on it. In modeling human cognition, there's always much left to do.

Acknowledgments

The first author was supported in part by ONR grant N00014-98-1-0332. The authors wish to acknowledge essential contributions from the Conscious Software Research Group: Ashraf Anwar, Ramesh Aitipamula , Arpad Kelemen, Ravikumar Kondadadi, Irina Makkaveeva , Lee McCauley, Aregahegn Negatu, Uma Ramamurthy, Alexei Stoliartchouk, and Zhaohua Zhang.

References

Allen, J. J. 1995. *Natural Language Understanding*. Redwood City CA: Benjamin/Cummings.

Anderson, J. R. 1990. *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Baars, B. J. 1988. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Barsalou, L. W. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences* 22:577–609.

Bogner, M., U. Ramamurthy, and S. Franklin. 2000. "Consciousness" and Conceptual Learning in a Socially Situated Agent. In *Human Cognition and Social Agent Technology*, ed. K. Dautenhahn. Amsterdam: John Benjamins.

Damasio, A. R. 1994. *Descartes' Error*. New York: Gosset; Putnam Press.

Edelman, G. M. 1987. *Neural Darwinism*. New York: Basic Books.

Ericsson, K. A., and W. Kintsch. 1995. Long-term working memory. *Psychological Review* 102:21–245.

Franklin, S. 1997. Autonomous Agents as Embodied AI. *Cybernetics and Systems* 28:499–520.

Franklin, S. 2000. Learning in "Conscious" Software Agents. In *Workshop on Development and Learning*. Michigan State U.; East Lansing, USA: April 5-7.

Franklin, S., and A. C. Graesser. 1997. Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. In *Intelligent Agents III*. Berlin: Springer.

Franklin, S., and A. Graesser. 1999. A Software Agent Model of Consciousness. *Consciousness and Cognition* 8:285–305.

Franklin, S., A. Kelemen, and L. McCauley. 1998. IDA: A Cognitive Agent Architecture. In *IEEE Conf on Systems, Man and Cybernetics*. : IEEE Press.

Hirst, T., and T. Kalus; 1998. Soar Agents for OOTW Mission Simulation. 4th Int'l Command and Control Research and Technology Symposium. September.

Hofstadter, D. R., and M. Mitchell. 1994. The Copycat Project. In *Advances in connectionist and neural computation theory, Vol. 2: logical connections*, ed. K. J. Holyoak, & J. A. Barnden. Norwood N.J.: Ablex.

James, W. 1890. *The Principles of Psychology*. Cambridge, MA: Harvard University Press.

Jurafsky, D., and J. H. Martin. 2000. *Speech and Language Processing*. Englewood Cliffs, NJ: Prentice-Hall.

Kanerva, P. 1988. *Sparse Distributed Memory*. Cambridge MA: The MIT Press.

Kintsch, W. 1998. *Comprehension*. Cambridge: Cambridge University Press.

Laird, E. J., A. Newell, and Rosenbloom P. S. 1987. SOAR: An Architecture for General Intelligence. *Artificial Intelligence* 33:1–64.

Libet, B., C. A. Gleason, E. W. Wright, and D. K. Pearl. 1983. Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential). *Brain* 106:623–642.

Maes, P. 1989. How to do the right thing. *Connection Science* 1:291–323.

Maturana, H. R. 1975. The Organization of the Living. *International Journal of Man-Machine Studies* 7:313–332.

McClelland, J. L., et al. 1986. *Parallel Distributed Processing*, vol. 1. Cambridge: MIT Press.

Ramamurthy, U., S. Franklin, and A. Negatu. 1998. Learning Concepts in Software Agents. In *From animals to animats 5*, ed. R. Pfeifer, et al Cambridge,Mass: MIT Press.

Sloman, A. 1987. Motives Mechanisms Emotions. *Cognition and Emotion* 1:217–234.

Sloman, A. 1999. What Sort of Architecture is Required for a Human-like Agent? In *Foundations of Rational Agency*, ed. M. Wooldridge, and A. Rao. Dordrecht, Netherlands: Kluwer Academic Publishers.

Zhang, Z., S. Franklin, B. Olde, Y. Wan, and A. Graesser. 1998b. Natural Language Sensing for Autonomous Agents. In *Proceedings of IEEE International Joint Symposia on Intellgence Systems 98*.

Reversing Category Exclusivities in Infant Perceptual Categorization: Simulations and Data

Robert M. French, Martial Mermillod (rfrench, mmermillod@ulg.ac.be)
Psychology Department, Université de Liège, Belgium

Paul C. Quinn (pquinn@washjeff.edu)
Psychology Department, Washington & Jefferson College, Washington, PA 15301 USA

Denis Mareschal (d.mareschal@bbk.ac.uk)
Centre for Brain & Cognitive Development, School of Psychology, Birkbeck College, London, U.K

Abstract

Three- to four-month-old infants presented with a series of cat or dog photographs show an unusual asymmetry in the exclusivity of the perceptual category representations formed. We have previously accounted for this asymmetry in terms of an inclusion asymmetry in the distribution of features present in the cat and dog images used during familiarization (Mareschal, French, & Quinn, 2000). We use a combination of connectionist modeling and experimental testing of infants to show that the asymmetry can be reversed by an appropriate pre-selection and minor image modification of cat and dog exemplars used for familiarization. The reversal of the asymmetry adds weight to the feature distribution explanation put forward by Mareschal et al. (2000).

Introduction

The ability to categorize is, without question, one of the central pillars of cognition. It is, therefore, not surprising that categorization abilities are present in humans from the very earliest age. Indeed, infants only a few months old are able to separate complex visual stimuli into generic object categories (e.g., Quinn & Eimas, 1996). In previous work, we have presented a simple connectionist model of perceptual categorization during early infancy (Mareschal & French, 1997; Mareschal & French, 2000; Mareschal et al., 2000). The model provided a mechanistic account of early infant category learning in terms of the data compression properties of connectionist autoencoder networks. Not only did this model capture standard infant categorization phenomena such as prototype formation and the use of feature co-variation information to form categories (Mareschal & French, 2000), but it also captured some of the more subtle idiosyncratic characteristics of infants' categorization behavior.

In particular, 3- to 4-month-olds show an unexpected asymmetry in the exclusivity of the perceptual category representations formed for cats versus dogs (Quinn, Eimas, & Rosenkrantz, 1993; Eimas, Quinn, & Cowan, 1994). Following exposure to a series of cat photographs, these infants will form a perceptual representation for cats that excludes dogs. In contrast, following exposure to a series of dog photographs, the same infants will form a category representation for dogs that does NOT exclude cats. Thus, there is an asymmetry in the exclusivity of the cat and dog categories: dogs are excluded from the representation for cats, whereas dog do not exclude cats.

We extend these results by showing how an opposite exclusivity asymmetry can be induced in 3- to 4-month-olds by a judicious choice of cat and dog exemplars presented to the infants prior to testing. Success in reversing the asymmetry between the Cat and Dog categories would lend strong support to a bottom-up account of early infant perceptual categorization.

Asymmetric exclusivity in infant categorization

Quinn et al. (1993) reported the following surprising categorization asymmetry. When 3- to 4-month-old infants were shown different photographs of either cats or dogs, they formed perceptual category representations for either groups of pictures. Infants were first shown a number of different photographs of cats and were then shown a picture of a dog paired with a novel picture of a cat. During the preference trials, the infants were much more attentive to the dog than to the novel cat. This was interpreted as showing that the infants had formed a category representation of Cat that excludes dogs. The dog, in other words, was perceived by the infants as not belonging to the category of cats. In sharp contrast, infants who were first shown a series of photographs of different dogs and were then shown a picture of a cat along with a novel dog were not preferentially attentive to either picture. When coupled with the finding that infants did not show a prior preference for cat photographs over dog photographs, and that infants familiarized with either cats or dogs looked longer at a bird photograph, the overall pattern of results was interpreted as showing that infants had formed a category representation of Dog that did not exclude cats. In short, the Dog category included cats, but the Cat category did not include dogs.

Infant perceptual categorization tasks frequently rely on preferential looking techniques based on the finding that infants direct attention more to unfamiliar or unexpected stimuli (e.g., Sokolov, 1963; Charlesworth, 1969; Cohen, 1973). While it is true that infants may sometimes have a preference for familiar stimuli, such as word stress patterns (Jusczyk, Cutler, & Redanz, 1993), it has been repeatedly shown that there is preferential attention directed to *novel* visual stimuli. The standard interpretation of this behavior is that the infants are comparing the input stimuli to an internal representation of those stimuli. As long as there is a discrepancy between the information stored in the internal representation and the visual input the infant continues to attend to the stimuli. While attending to the stimuli the infant updates its internal representation. When the information in the internal representation is no longer discrepant with the visual input, attention is switched elsewhere. When a familiar object is presented there is little or no attending because the infant already has a reliable internal representation of that object. In contrast, when an unfamiliar or unexpected object is presented, there is a lot of attending because an internal representation has to be constructed or adjusted.

When a series of exemplars can be grouped into a consistent category, this account of representation construction implies a progressive decrease in looking time with successive exemplars encountered. Although each exemplar encountered is novel (and therefore attracts the infant's attention), the process of representation construction gradually leads to the extraction of key dimensions of the category. Thus, after some time, a reliable category representation is constructed and new exemplars encountered (although still novel), take little time to be assimilated to the internal category representation and therefore only briefly capture the infant's attention.

A model of infant perceptual categorization

We used a three-layer autoencoder to model infant categorization behaviors (Mareschal & French, 1997; Mareschal & French, 2000; Mareschal et al., 2000). Learning in an autoencoder consists of developing an internal representation of the input (at the hidden unit level) that is sufficiently reliable to reproduce all the information in the original input (Cottrell, Munro, & Zipser, 1988). Information is first compressed into an internal representation and then expanded to reproduce the original input. The successive cycles of training in the autoencoder are an iterative process by which a reliable internal representation of the input is developed. The reliability of the internal representation is tested by expanding it and comparing the resulting predictions to the actual stimulus being encoded.

The degree to which network error increases on presentation of a novel object depends on the similarity between the novel object and previously seen (i.e., familiar) objects. Presenting a series of similar objects leads to a progressive drop in error on future similar objects. The modeling assumption that we have therefore made is that network error and infant attention levels correlate: the higher the network error, the longer the looking time of the infant This is true of both autoassociators (where output error is the measurable quantity) and infants (where looking time is the measurable quantity).

To model the Cat/Dog findings, we obtained data for the networks from measurements of the original cat and dog pictures used by Quinn et al. (1993). There were 18 dogs and 18 cats classified according to the following ten traits: head length, head width, eye separation, ear separation, ear length, nose length, nose width, leg length vertical extent, and horizontal extent. The feature values were normalized over all pictures in both training sets to be within 0 and 1.

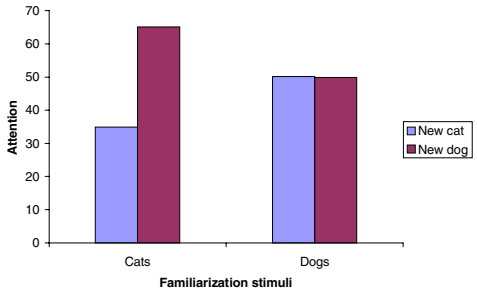


Figure 1. Generalization errors for networks trained on cats and dogs (Mareschal *et al*. 2000).

Figure 1 shows what happens when networks trained on cats are presented with a novel cat and a dog, and when networks trained on dogs are tested with a novel dog and a cat. When the networks are initially trained (i.e., familiarized) on cats, the presentation of a dog results in a large error score (corresponding to infants' longer looking time). Dogs are not included in the category representation of cats. In contrast, when the networks are initially trained on dogs, the presentation of a cat will result in essentially the same error as a novel dog, suggesting that the cats have been included in the category representation for dogs.

Because auotencoders extract the distribution statistics of the exemplars they have encountered, this led us to explore the distribution of feature values in the data measured from the original photographs in order to explain the asymmetry. Figure 2 shows the probability distributions of the 10 traits for both cats and dogs. Some of the traits are very similar in terms of their means and distribution of both cats and dogs (e.g., head length and head width). Others, especially nose length and nose width, are very different and will provide the crucial explanation of the unexpected attentional asymmetries reported by Quinn et al. (1993) and Eimas

et al. (1994). It is clear that in almost all cases the distribution for each Dog trait (represented by the dark line) subsumes the distribution for corresponding trait for cats. The narrower distributions for most Cat traits, on the other hand, do not subsume the range of values for the corresponding Dog traits. In other words, cats are possible dogs but the reverse is not the case: most dogs are not possible cats. Specifically, when we examine all of the members of the two populations, we see that the values of all ten traits for 9 (i.e., 50%) of the members of the Cat category fall within a 2σ cut-off for those traits for the Dog category. Fully half of the cats in the population could be reasonably classified as dogs. In contrast, the smaller means and lower variances of a number of traits (especially, nose length and nose width) for cats compared to dogs means that only 2 of the 18 dogs could conceivably be classified as being members of the Cat category.

Hence, it seems that the exclusivity asymmetry is driven by (1) an associative learning mechanism that is sensitive to feature distributions, and (2) a distribution

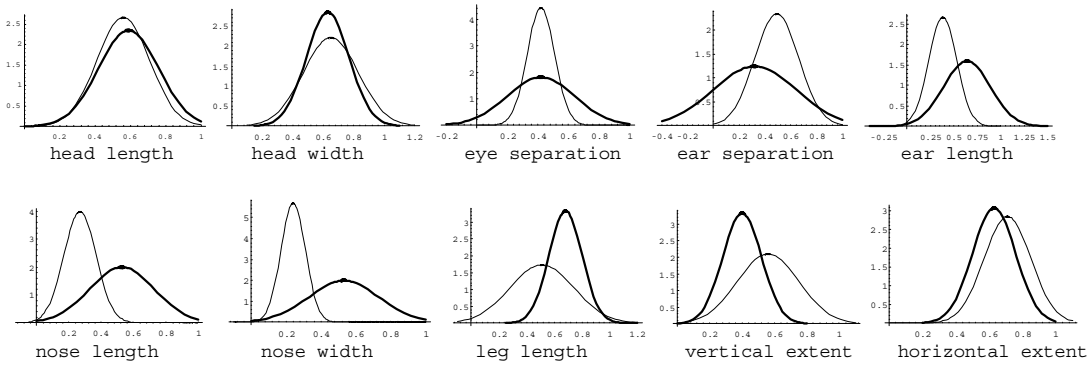


Figure 2. Frequency distributions for the ten defining traits of 18 dogs and 18 cats in Mareschal et al., 2000. The variance of Dog traits is, on average, 1.6 times that of Cats. Dogs' features largely subsume by cats'.

profile in which the Dog feature values largely subsume the Cat feature values. A direct implication of this is that if the distribution statistics were reversed, then we should observe a reversed categorization asymmetry. In this new case, infants should develop a perceptual category representation of Dog that excludes cats and a perceptual category representation of Cat that does not exclude dogs. The simulation and experiments reported below test this prediction directly.

Reversing asymmetric exclusivity

To explore whether the asymmetry could be experimentally reversed, we began by artificially manipulating the naturally occurring variance of the two categories. In the original experiment the within-category variability of the dog photographs was greater than that of the cat photographs and, crucially, the feature set for dogs largely subsumed that of cats. However, by carefully selecting sets of cat and dog photographs and then morphing a number of these images, we were able to reverse the variance of the categories In the original experiment (Mareschal et al., 2000) the average variance over all ten features of the Dogs was 1.63 times that of the Cats, whereas for the modified images the average variance of Cats was 3.12 times that of Dogs. Figure 3 shows the feature distributions for these modified exemplar sets. A comparison with the original data plotted in Fig. 2 shows how the distributions have been reversed. This is especially clear for features such as "Leg length." There were an identical number of morphed images (8 out of 18) in both the Dog and Cat stimuli sets.

Reversing the network's learning

The simulation reported was done on a standard 10-8-10 feedforward backpropagation autoencoder network (learning rate: 0.1, momentum: 0.9, Fahlman offset: 0.1). Training was identical to that in Mareschal et. al. (2000). Networks were trained in batches of 2 patterns for a maximum of 250 epochs. This simulated familiarization with pairs of pictures for a fixed period before being presented with a new familiarization pair. Results were averaged over 50 runs.

Figure 4 shows the model's generalization error to novel exemplars of cats and dogs as a function of whether they were trained on cats or on dogs. Networks trained with cats show no difference in error (hence predict no difference in looking times) when tested with a novel cat or a dog. In contrast, the networks originally trained with dogs show much greater error when tested with a novel cat than when tested with a novel dog (suggesting a strong preference for looking at a cat vs. a novel dog). This asymmetry is exactly the opposite of the one found in the original Mareschal et al. (2000) study and constitutes an explicit prediction of the autoencoder model.

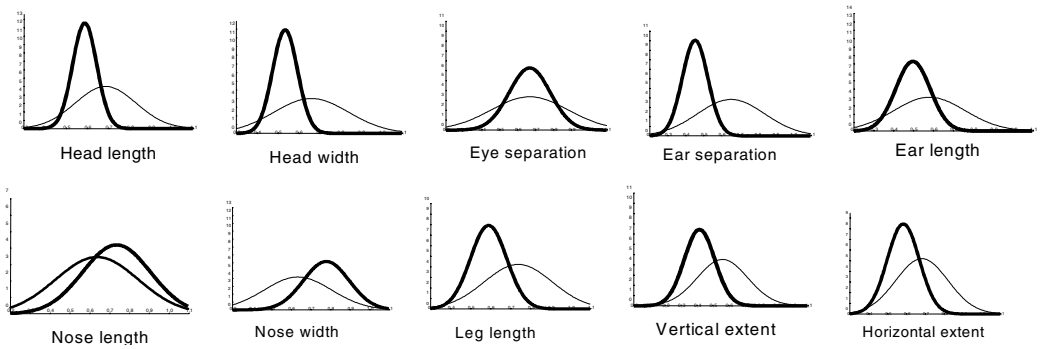


Figure 3. Frequency distributions for Exp. 1 for the 18 dogs and 18 cats. The variance has been artificially reversed and, crucially, Dog features are largely subsumed by Cat features. Compare with Figure 2

Experiment 1: Reversal of exclusivity

This prediction was tested with two groups of 3- and 4-month-old infants that were presented with a set of 12 exemplars from the same category, cats or dogs, during a series of familiarization trials, and were then presented with preference test trials consisting of a novel cat paired with a novel dog. The model predicts that infants familiarized with dogs should display a novel category preference for cats on the preference test trials, whereas infants familiarized with cats should display looking times divided evenly between the dogs and cats displayed on the preference test trials.

Method

Participants. The participants in Exp. 1 were 12 infants approximating 3 and 4 months of age (M = 3 months, 20 days; SD = 8.30 days). Seven of the infants were females and five were males.

Stimuli. The stimuli were 36 colored photographs of cats and dogs (18 exemplars per category, representing 18 different breeds for each category). The photographs were obtained from Siegal (1983) and Schuler (1980). In order to obtain dogs with low perceptual variance and cats with high variance, certain stimuli were slightly modified using computer imaging processing software (Rubber v.2.0). None of the stimuli were "morphed" to the point of giving the impression of a strange animal. The same number of animals (8) were morphed in both groups. The variance of the Cat category was modified so that the average variance of Cats was 3.1 times that of Dogs (compared to the original experiment where the average variance of Dogs was 1.6 times that of Cats). As in Quinn et al. (1993) and Mareschal et al. (2000), the pictures selected were chosen to represent a variety of shapes, colors, and orientations of each type of animal. The size of the animal in each picture was nearly the same, and thus not a reflection of its actual size (so that any categorization effects observed would not be the result of simple size discrimination.)

Each stimulus contained a single animal, cut away from its background, centered, and mounted onto a white 17.7 x 17.7 cm posterboard for presentation.

Apparatus. Infants were tested in a visual preference apparatus, modeled on the one described by Fagan (1970). The apparatus is a large, three-sided gray viewing chamber that is on wheels. It has a hinged, gray display panel (85 cm high and 29 cm wide) onto which were attached two compartments to hold the posterboard stimuli. The stimuli were illuminated by a 60-Hz fluorescent lamp that was shielded from the infant's view. The center-to-center distance between compartments was 30.5 cm and on all trials the display panel was situated approximately 30.5 cm in front of the infant's face. There was a 0.625 cm peephole midway between the two display compartments allowing observation of the infant's visual fixations.

Procedure. In both experiments, infants were placed in a reclining position on their seated parent's lap. An experimenter positioned the apparatus such that the midline of the infant's head was aligned with the midline of the display panel. The experimenter loaded the appropriate stimuli into the display panel, elicited the infant's attention and exposed the stimuli to the infant. During the course of a trial, the experimenter observed the infant through the peephole and recorded visual fixations to the left and right stimuli by means of two 605 XE Accusplit electronic stop watches, one of which was held in each hand. Interobserver reliability for this procedure was determined by comparing the looking times measured by the experimenter using the center peephole and additional observers using peepholes to the left of the left stimulus compartment and to the right of the right stimulus compartment is high (Pearson r = .97), a value comparable to values obtained in other laboratories that measured visual fixation duration with the corneal reflection procedure (e.g., O'Neil, Jacobson, & Jacobson, 1994). Two experimenters recorded fixations, one during familiarization and another during preference test trials. Importantly, the person recording during preference test trials was unaware of the category information that was presented during the familiarization trials.

Each infant was assigned twelve randomly selected

pictures of cats or dogs. On each of six 15s familiarization trials, two of the twelve stimuli, again randomly selected, were presented. Six infants were randomly assigned to each group, defined by the familiarization category, cats or dogs. Immediately after the familiarization trials, two 10s preference test trials were administered in which a novel cat was presented with a novel dog. There were six such pairs, randomly selected, and each pair, which was seen on both test trials, was assigned to one infant who had seen dogs and one infant who had seen cats during the familiarization trials. The test-trial stimuli were thus identical for both groups of infants. The left-right positioning of the novel animal from the novel category was appropriately counterbalanced across infants.

	Familiarization phase (average fixation time in secs.)		Novelty preference (% of viewing time for unfamiliar category)	t
Familiarization category	First 3 trials	Last 3 trials		
Cats	7.8(3.8)	6.9(3.6)	49.5% (16.7)	-0.08
Dogs	7.9(1.4)	9.2(3.1)	70.4% (9.7)	5.1*

t for mean vs. chance *p < .005, one-tailed.

Table 1. Mean fixation times in Experiment 1.

Results and Discussion

Familiarization trials. Individual looking times were summed over the two stimuli that were presented on each trial and then averaged across the first three and the last three trials. The mean looking times and standard deviations are shown in Table 1. Novelty preference is expressed in percentage of time that the infant looks at the exemplar from the unfamiliar category compared to the total time regarding the pair of test stimuli. An analysis of variance, familiar category (cats vs. dogs) x trial block (1-3 vs. 4-6), performed on the individual scores, revealed no significant effects, $F(1, 10) < 2.28$, $p > .15$, in each instance. As has been the case in previous infant categorization studies using the same procedures and similar stimuli (Eimas & Quinn, 1994; Eimas, et al. 1994; Mareschal et al., 2000; Quinn & Eimas, 1996, 1998; Quinn, et al., 1993), no evidence of habituation was obtained. We believe the complexity and variety of stimuli were sufficient to maintain infant attention during the familiarization trials.

Preference test trials.
The total looking time of each infant across the two test trials to the novel stimulus from the novel category was divided by the total looking time to both test stimuli and converted to a percentage score. Mean novel category preference scores are shown in Table 1 and in Figure 5. t tests versus chance, which were one-tailed because of the predicted preference in the direction of novelty, revealed that infants familiarized with dogs preferred the novel cat, but infants familiarized with cats did not prefer the novel dog. In addition, the two means were significantly different from each other, $t(10) = 2.65$, $p < .05$, two-tailed. As was predicted by the model, infants familiarized with dogs formed a category representation that excluded cats, but infants familiarized with cats did not form a category representation that excluded dogs. The findings are exactly the opposite of those reported in Exp. 4 of Quinn et al. (1993). Thus, we can reasonably conclude that the stimulus manipulations were successful in reversing the inclusion relation between dogs and cats reported by Mareschal et al. (2000).

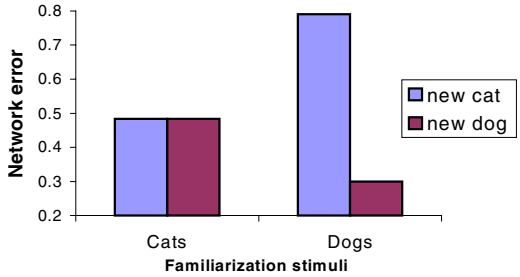


Figure 4. Network generalization errors when Cat features largely subsume Dog features

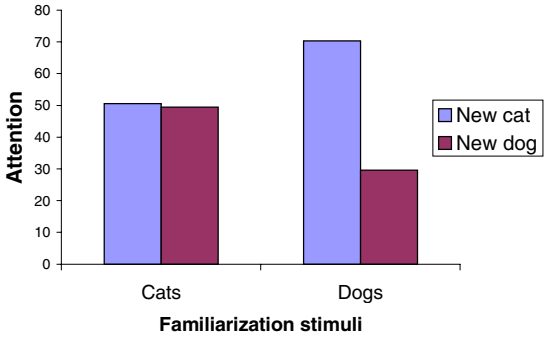


Figure 5. Infant attention when Cat feature distributions largely subsume those for Dogs.

Experiment 2: No prior preference

An alternative explanation for the outcome of Exp. 1 is that infants might have spontaneously preferred the cats over the dogs. Although no such spontaneous preference was found in Quinn et al. (1993) or Quinn and Eimas (1996), it is possible that the stimulus manipulations could have inadvertently produced one. If there was a preference for cats over dogs in Exp. 1, then it would have facilitated (if not fully explained) any presumed novel category preference for cats after familiarization with dogs, and it would have interfered with any novel category preference for dogs after familiarization with cats. Exp. 2 was thus replication of Exp. 1, but conducted without the familiarization trials.

Method

Participants. 6 infants approximately 3 and 4 months old (M = 3 months, 17 days; SD = 12.71 days). Three of the infants were females and three were males.

Stimuli and apparatus. Same as Exp. 1.

Procedure. All infants were presented with the preference test trials described for Exp. 1, but without the prior familiarization trials.

Results and Discussion

A preference score for cats was determined for each infant for the two trials by dividing the looking time that the cat was observed by the total looking time to both the cat and dog. The score was then converted into a percentage. The mean preference for cats was 48.34%, SD =22.03. This preference was not significantly different from chance, $t(5) = -0.18$, $p > .20$, two-tailed. Further, the preference for cats after familiarization with dogs in Exp. 1 was found to be reliably higher than the spontaneous preference for cats with no familiarization with dogs, $t(10) = 2.24$, $p < .05$, two-tailed. The preference results from Exp. 1 are thus unlikely to be reflective of a spontaneous preference for cats and more likely are a consequence of the reversal of the inclusion relation between cats and dogs.

Conclusion

Quinn, Eimas, & Rosenkrantz (1993) observed a striking asymmetry in the infant categorization of photos of cats and dogs. An initial simulation by Mareschal and French (1997) and Mareschal et al. (2000) was able to reproduce the original experimental results by focusing on the *within-category variability and inclusion relation* of the two categories of animals. This simulation led to a prediction — namely, that if the degree of variability and overlap of shared feature distributions was the key to explaining this categorization asymmetry, then artificially reversing the order of the within-category variability for shared features should reverse the infant categorization asymmetry. We were able to reverse this categorical variability and, as predicted by the model, we observed the reverse categorization asymmetry in the infants.

The reversal of the asymmetry makes the point that infants form at least some category representations on-line, rather than tapping into pre-existing concepts that had been formed prior to arriving at the laboratory. If the infants been relying on previously acquired categories, then their responsiveness should not have varied with the variations in the familiar category information presented. The fact that infant responsiveness did vary across experiments suggests that the categories were being formed on-line and that the boundaries can be pushed around depending on the information presented during familiarization.

Acknowledgments

This work was supported by Grants BCS-0096300 from the NSF to P. Quinn, RS000239112 from the ESRC (UK) to D. Mareschal and HPRN-CT-2000-00065 from the Eur. Commission to R. French. Thanks to C. Labiouse for his comments on this work and to L. Yarzab for her assistance with the experiments.

References

Charlesworth, W. (1969). The role of surprise in cognitive development. In D. Elkind & J. Flavell (Eds.), *Studies in cognitive development. Essays in honor of Jean Piaget*, pp. 257-314, Oxford University Press.

Cohen, L. (1973). A two-process model of infant visual attention. *Merrill-Palmer Quarterly, 19,* 157-180.

Cottrell, G., Munro, P., & Zipser, D. (1988). Image compression by backpropagation: and example of extensional programming. In N. E. Sharkey (Ed.), *Advances in cognitive science*, *Vol. 3*. Ablex.

Eimas, P., & Quinn, P. (1994). Studies on the formation of perceptually based basic-level categories in young infants. *Child Development, 65,* 903-917.

Eimas, P., Quinn, P., & Cowan, P. (1994). Development of exclusivity in perceptually-based categories of young infants. *J. of Exp. Child Psychology, 58,* 418-431.

Fagan, J. (1970). Memory in the infant. *J. of Experimental Child Psychology, 9*, 217-226.

Jusczyk, P., Cutler, A., & Redanz, N. (1993). Infants' preference for the predominant stress patterns of English verbs. *Child Development, 64,* 675-687.

Mareschal, D. & French, R. (2000). Mechanism of categorization in infancy. *Infancy, 1*, 59-76.

Mareschal, D. & French, R. (1997). A connectionist account of interference effects in early infant memory and categorization. *Proceedings of the 19th Annual Cognitive Science Society Conference*, LEA, 484-489.

Mareschal, D., French, R., & Quinn, P. (2000). A connectionist account of asymmetric category learning in early infancy. *Developmental Psych., 36,* 635-645.

O'Neill, J., Jacobson, S., & Jacobson, J. (1994). Evidence of observer reliability for the Fagan Test of Infant Intelligence. *Infant Behavior & Dev, 17,* 465-469.

Quinn, P. & Eimas, P. (1996). Perceptual cues that permit categorical differentiation of animal species by infants. *J. of Experimental Child Psychology, 63,* 189-211.

Quinn, P. & Eimas, P. (1998). Evidence for a global categorical representation of humans by young infants. *J. of Experimental Child Psychology, 69,* 151-174.

Quinn, P., Eimas, P., & Rosenkrantz, S. (1993). Evidence for representations of perceptually similar natural categories by 3- and 4-month-old infants. *Perception, 22,* 463-475.

Schuler, E. M. (Ed.) (1980). *Simon and Schuster's guide to dogs.* New York: Simon and Schuster.

Siegal, M. (Ed.) (1983). *Simon and Schuster's guide to cats.* New York: Simon and Schuster.

Sokolov, E. N. (1963). *Perception and the conditioned reflex.* Hillsdale, NJ: LEA.

Adaptive Selection of Problem Solving Strategies

Danilo Fum * (fum@univ.trieste.it)
Department of Psychology, via S. Anastasio 12
Trieste, I-34134 Italy

Fabio Del Missier (delmisfa@univ.trieste.it)
Department of Psychology, via S. Anastasio 12
Trieste, I-34134 Italy

Abstract

The issue of strategy selection in solving the Tower of Hanoi (TOH) problem is investigated by focusing on the critical issues of whether the selection process is contingent and adaptive. The results of an experiment in which participants solved a series of different four-disk TOH problems under instructions requiring accuracy maximization vs. effort minimization are presented. A computer simulation, comparing a number of known strategies to the experimental data, has been carried out to try to identify the strategies used by the participants. The findings support the hypothesis of adaptive and contingent strategy selection in the TOH domain.

Introduction

Much work in the problem solving arena has dealt with the Tower of Hanoi (TOH) – considered as a typical well-structured problem – producing important theoretical and empirical results. Researchers have discovered interesting phenomena and tried to provide explanations for them. Several solution strategies have been described (Simon, 1975), and various models have been proposed to simulate human performance on this task (Karat, 1982; Ruiz & Newell, 1989; Anderson, Kushmerick & Lebiere, 1993; Anderson & Lebiere, 1998; Altmann & Trafton, 2000). Detailed accounts of learning how to solve the TOH on a trial-by-trial basis (Anzai & Simon, 1979) have been put forward together with hypotheses concerning the strategies and the heuristics people seem to learn in successive attempts to solve the problem (VanLehn, 1991).

Despite these achievements, many issues are still unresolved and many topics are currently investigated. Two recent examples involve the role of goal encoding and retrieving in memory (Altmann & Trafton, 2000), and the possible use of active planning to avoid previously visited states (Davies, 2000).

Given that different models and strategies have been proposed in different experimental settings, it seems important to try to identify the factors affecting the selection of solution strategies in this domain.

We propose the hypothesis that strategy selection in the TOH is a contingent processes, i.e., it is sensitive to task and contextual factors. Following a widely accepted idea about human problem solving (Simon, 1975; Anderson, 1990; Christensen-Szalanski, 1998) and decision-making (Payne, Bettman & Johnson, 1993), it is further hypothesized that strategy selection is adaptive. Given a specific task and context, it is functional to the achievement of a good trade-off between accuracy and cognitive effort (Christensen-Szalanski, 1998; Fum & Del Missier, 2000).

These two strategy-related questions (i.e., is the process of strategy selection contingent? is it adaptive?) are the main topics of this work. In the paper we briefly discuss some issues concerning research on the TOH strategies. Then we present the results of an experiment in which participants solved a series of different four-disk TOH problems under instructions requiring accuracy maximization vs. effort minimization. A computer simulation, comparing several solution strategies to the experimental data, has been carried out to try to identify the strategies used by participants in the two instruction groups.

Issues on Strategy Research

Research on strategies in TOH, and related problem solving tasks, must deal with several theoretical and empirical issues.

A first issue concerns identificability (Anderson, 1990): patterns of behavioral data are used as a trace to induce the existence of a given strategy, but in many cases the data do not allow discriminating among distinct models of strategic behavior. In our specific domain, however, very few attempts (an exception being represented by Altmann & Trafton, 2000) of directly comparing different models on the same data set have been done.

Other theoretical problems deal with the under-specification and the low generalizability of some of the proposed strategies. With underspecification we mean

*The order of authorship is arbitrary; each author contributed equally to all phases of this project.

the fact that the description of a strategy does not allow a unique identification of the move to be done for every problem state. With low generalizability we mean the fact that the proposed strategy results ad hoc and cannot be extended to deal with some classes of TOH problems people are able to solve.

A further theoretical limitation is constituted by the fact that some strategies are willfully optimal (Anderson & Lebiere, 1998), while people seldom achieve such a brilliant performance (Goel & Grafman, 1995; Miyake et al., 2000; Karat, 1982).

On the empirical side, there is the problem of the intrusiveness of the methods utilized to identify the existence of a given strategy. Verbal protocols, for instance, (Anzai & Simon, 1979; Van Lehn, 1991) have proved to be a useful exploratory tool, but there is evidence (Stinessen, 1985; Ahlum-Heath & DiVesta, 1986) that participants verbalizing during the task perform differently from participants that do not verbalize. The very use of verbal protocols could prompt the adoption of different solution strategies.

A related issue deals with the suggestiveness of the experimental instructions. For instance, Anderson, Kushmerick & Lebiere (1993) gave hints that deliberately encouraged the adoption of a particular strategy. The generalizability of their model is, therefore, directly related to the way the same strategy is spontaneously adopted by the participants when no hints are given.

Another concern is constituted by the fact that strategy selection in the TOH has often been studied by having people perform many trials over the same problem. In this way it cannot be excluded that the improvement in the participants performance could be attribute to rote memorization instead of genuine learning. To control for this factor, Anderson, Kushmerick & Lebiere (1993) presented a wider range of problems to their participants preventing them from evolving special-case strategies.

In our experiment we investigated a factor that could possibly affect the adoption of different solution strategies, and we ran a simulation study to try to identify them. To do this, we had to make some underspecified strategies computationally workable by postulating a few additional assumptions. We concentrated our attention on general strategies – i.e. on strategies capable of solving problems put not only in their standard (i.e., tower-to-tower) form – and on strategies that do not prescribe an optimal solution. Furthermore, we refrained to force participants to justify and comment on their moves, and carefully avoided suggesting any specific solution procedure. Finally, we utilized a set of different problem types.

The Experiment

The main goal of the experiment was to test the hypotheses of contingent and adaptive strategy selection. We manipulated the experimental instructions to modify the importance participants gave to the distinct goals of accuracy maximization vs. effort minimization.

According to the contingent and adaptive hypothesis, we expected to find a rational use of different strategies in different experimental groups. The strategies used by participants in the accuracy group should increase the accuracy of the solutions by paying a higher temporal cost. The strategies used in the effort group should yield effort savings but less accurate solutions.

Method

Participants The participants were 34 undergraduates students, aged between 18 and 24. None of them was suffering from any perceptual, cognitive or motor deficiency. The sample was balanced for gender. All the participants had a basic familiarity with computers and were able to use the mouse.

Procedure Participants read an instruction document that explained the basic rules of the TOH, showed the interface used by the computer program, and described how to use it. The instructions required the participants to solve the problem "in the fewest possible number of moves" or "in the shortest possible time", depending on the group (accuracy vs. effort, respectively) to which they were randomly assigned. The experimenter (always one of the authors) asked the participants about their knowledge of the task and was willing to answer possible questions about the procedure. After going through a short training session, participants started to solve the series of test problems.

Materials A number of different three- and four-disk TOH problems were randomly generated for the experiment. The problems comprised four possible configurations of disks obtained by combining a flat vs. tower disposition in the start state with a flat vs. tower disposition in the goal state.

Two randomly generated three-disk problems, with an optimal solution path of seven moves and with a flat-to-flat configuration, were used for training and presented to the participants in casual order.

The test set comprised eight randomly generated four-disk problems, two for each possible configuration. Each problem had an optimal solution path of 15 moves. The test set was delivered using block randomization.

Apparatus A PowerMacintosh 9500 computer was used for the experiment. A program implementing the

TOH task was written using MCL 4.3 and CLIM 2. The program recorded each participant move (including the moves violating the TOH rules) with the associated time.

The interface was composed by two identical windows, vertically stacked and centered. The upper window showed the initial state of the problem and could be acted upon by the participants. The lower window, which showed the goal state, presented a fixed display. The participants had to perform a drag-and-drop operation with the mouse to move disks from peg to peg in the upper window. In case of an illegal move, an auditory warning was delivered, and the dragged disk was forced back to its source peg.

Experimental Design Two independent variables – one between-subjects (instruction type) and one within-subjects (trial number) – were manipulated in a 2x8 mixed design. The number of trials in the test session (eight) was chosen to obtain an acceptable balance between the possibility of obtaining learning effects and that of inducing fatigue effects. The basic dependent variables were the number of errors (i.e. illegal moves in addition to minimum path length), the number of attempted illegal moves, the total time to solve the problem, the mean move latency (excluding the first move), and the time necessary to execute the first move.

Results

All the data analyses were performed on 31 cases[1] (15 in the accuracy, 16 in the effort group) either on transformed and untransformed variables[2]. Given the absence of any difference, we will present only the results obtained using the untransformed variables.

Errors A 2x8 analysis of variance (ANOVA) on the number of errors (Figure 1) showed the significant main effects of instruction type ($F(1,29)=6.57$, $MSE=173.53$, $p<.05$), and trial ($F(7,203)=4.95$, $MSE=69.33$, $p<.001$). The interaction was not significant. The participants in the accuracy group made fewer errors that those in the effort group ($M=6.73$ for accuracy; $M=11.02$ for effort). In both groups the number of errors decreased from the first block of four trials to the second block ($M=11.38$ for the first block, $M=6.38$ for the second one). A post hoc analysis carried out with the Tukey HSD test showed significant differences between the following pairs of trials: 1-5 ($p<.05$), 1-6 ($p<.01$), 1-8 ($p<.05$), 2-5 ($p<.01$), 2-6 ($p<.001$) and 2-8 ($p<.01$). The Bonferroni procedure confirmed the results.

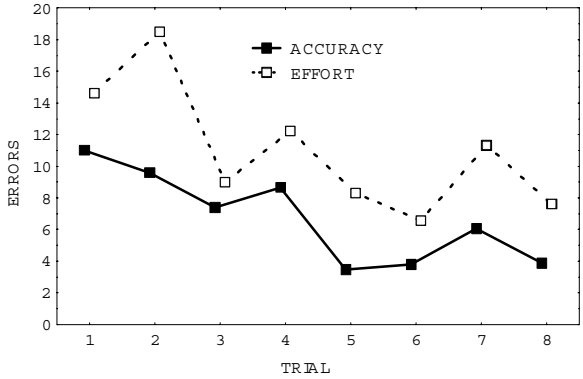


Figure 1: Number of errors for each trial in the accuracy and effort conditions.

Illegal Moves Participants attempted to execute very few illegal moves. The number of such moves was however lower in the accuracy group than in the effort group ($M=0.77$ for accuracy, $M=1.87$ for effort), and decreased from the first to the second block of trials ($M=2.00$ for the first; $M=0.65$ for the second block). Both the effects, but not the interaction, were statistically significant ($F(1,29)=6.71$, $MSE=11.17$, $p<.05$ and $F(7,203)=6.37$, $MSE=3.76$, $p<.001$, respectively).

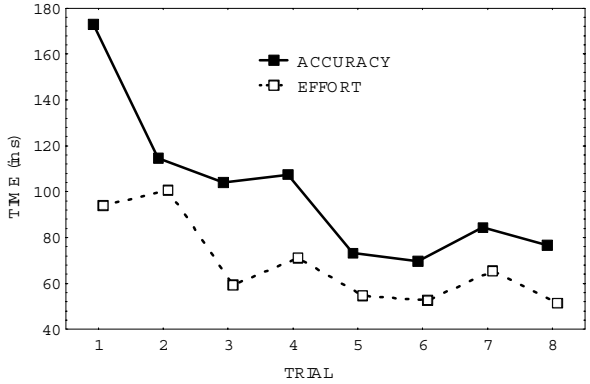


Figure 2: Solution times for each trial in the accuracy and effort conditions.

Solution Times A 2x8 ANOVA on the solution times (Figure 2) revealed the significant main effects of the instruction type ($F(1,29)=7.83$, $MSE=7947.67$, $p<.01$) and of the trial ($F(7,203)=8.68$, $MSE=2259.27$, $p<.001$), while their interaction was not significant. The participants needed more time to solve the problems in the accuracy than in the effort group ($M=100$ s for accuracy, $M=69$ s for effort). The time necessary to complete the task decreased from the first to the second block of trials ($M=103$ s for the first, $M=66$ s for the

[1] Two cases were excluded because the participants needed more than the maximum allowed time (45 min) to complete the first two problems in the test session. One case was excluded because the participant said, only at the end of the session, that she had previously written a program capable of solving this kind of task.

[2] A logarithmic transformation was performed on all the variables measuring time, while a square-root transformation was applied to all the variables recording the number of moves.

second). The Tukey test and the Bonferroni procedure highlighted significant differences between the first trial and the last six and between the second trial and the last four (with the exception of the pair 2-7).

Move Latencies A 2x8 ANOVA on move latency times showed a significant interaction ($F(7,203)=3.10$, $MSE = 946783$, $p< .01$) between instructions and trial.

The main effects were also significant: $F(1,29)=14.85$, $MSE=9131212$, $p< .001$ for instructions, $F(7,203)=8.00$, $MSE=946783$, $p< .001$ for the trial. A 2x7 ANOVA with the exclusion of the first trial confirmed the main effects but not the interaction ($F(6,174)=1.29$, $MSE=745583$, $p=.26$). This result suggests that the interaction could be attributed to the extremely high latencies of the participants in the accuracy group on the first trial. This was confirmed by the post hoc tests on the first ANOVA. The move latency was higher in the accuracy group than in the effort group ($M=3.82$ s for accuracy, $M=2.34$ s for effort), and decreased from the first to the second ($M=3.46$ s first block; $M=2.71$ s second block). The Tukey post hoc analysis on the second ANOVA showed significant differences between the pairs 2-5 ($p< .05$), 2-6 ($p< .05$) and 2-8 ($p< .05$). The Bonferroni procedure confirmed only the difference between the trials 2 and 8 ($p< .05$).

First Move Latency A 2x8 ANOVA on the first move latency showed only the significant main effect of the instruction type ($F(1,29)= 13.18$, $MSE=583.18$, $p< .01$) with latency higher for participants in the accuracy group ($M=14.78$ s for accuracy, $M=3.64$ s for effort).

Cluster Analysis of Move Latencies We performed also a k-means cluster analysis to determine whether the means of the move latencies and the mean percentages of moves within given latency boundaries were different between the two instruction groups. The cluster analysis was performed on all the moves that required less than 4 s to be executed[3]. For each subject a solution with 2 clusters (moves having an almost exclusive motor component vs. moves requiring more relevant cognitive processes) was looked for.

A 2x2 (Move x Instruction) ANOVA on the cluster means showed a significant interaction ($F(1,29)=9.46$, $MSE=80919$, $p< .01$) and significant main effects of the move kind ($F(1,29)=2499.81$, $MSE=10229$, $p< .001$) and of the instruction type ($F(1,29)=10.31$, $MSE=10229$, $p< .01$). The interaction is explained by the fact that the difference of 140 ms between participants in the two instruction groups for the "cognitive" moves ($M=3.08$ s for accuracy and $M=2.94$ for effort) was significantly smaller than the difference of 305 ms found between the groups for the simplest "execution" moves ($M=1.87$ for accuracy vs. $M=1.57$ for effort). These results confirm the indications obtained from the previous move latency analysis, but suggest also a potential execution speed-up for the participants in the effort group.

A further analysis was focused on the mean percentages of cases belonging to the two move clusters and to the moves requiring 4 s or more (the third cluster of "long" moves) in both instruction groups. The results showed significant differences between the accuracy and effort groups for the execution moves (Mann-Whitney U test, $U=57$, $z=2.49$, $p< .05$) and long ones ($U=32$, $z=3.47851$, $p< .001$). In particular, the mean percentage of cases belonging to execution moves was greater in the effort group ($M=61.77$, $SD=8.35$) than in the accuracy group ($M=50.94$, $SD=12.11$). The reverse was true for the long moves (accuracy: $M=24.82$, $SD=10.26$; effort: $M=12.08$, $SD=5.98$). This could mean that participants in the effort group made a higher percentage of execution moves and a lower percentage of cognitive moves in comparison with the moves made by the participants in the accuracy group.

Discussion

There is clear evidence that the experimental manipulation has been very effective in changing the way the TOH problems are solved. As expected, participants are able to achieve their respective goals of minimizing effort and maximizing accuracy, and they are forced by the instructions to trade a lower number of moves with a higher solution time.

There is also clear evidence of the existence of a learning effect. Participants in both groups learn to perform better in successive trials, making fewer errors and using less time. The learning profiles for the two groups remain however distinct across all the trials. The difference concerns not only the errors made and the times needed for solution, but extends to all the dependent variables suggesting that participants in the two groups were selecting and using different solution strategies.

The Simulation

The goal of the simulation was to try to identify the strategies used in each trial by participants in the two instruction groups by comparing several known TOH solution strategies on their capacity to fit the data.

[3] Given an independent estimate of 2.15 s for the time needed to move a disk using a TOH program with a direct-manipulation user interface (Anderson, & Lebiere, 1998), we assume that moves requiring 4 s or more are also affected by some kind of higher-order cognitive operation.

The Implemented Strategies

For the simulation we developed a series of ACT-R (Anderson & Lebiere, 1998) models implementing the following solution strategies:

SS1 The selective search strategy described by Anzai & Simon (1979), and subsequently studied by Van Lehn (1991). At each step only disks that are free to move in the current state are considered. The choice of which disk to move and where is guided by two heuristics: "(1) do not move the same disk on consecutive moves, and (2) do not move the smallest disk back to the peg it was on just before it was moved to its current peg" (Van Lehn, 1991, p. 6). Because the strategy is under-specified, an additional assumption has been made: "(3) whenever possible, choose the move which has the effect to put the largest out of place disk (the LOOP disk) into the target peg", which gives the strategy a more goal-oriented attitude. Because the participants did not always follow the directives of the don't-move-twice and don't-undo-move heuristics, the model employs them probabilistically according to two empirically-derived parameters (93% of the cases in which they could be applied when modeling the participants in the accuracy condition, and 90% of the times for the effort condition). Finally, whenever there is still uncertainty about which move to make, the model chooses randomly.

SS2 The selective search strategy previously described augmented with the new one-follows-two heuristics that states that if you have just moved the disk of dimension two, you should now put the smallest disk on top of it.

SP The (simple) perceptual strategy described in Simon (1975) and rephrased as follows: "(1) if all n disks are placed on the target peg, stop; else (2) find the next disk (i) to be placed on the target peg (3) if there are smaller disks on top of disk i, clear them (4) clear disks smaller than i off the target peg (5) move disk i to the target peg (6) go to 1." (Goel & Grafman, 1995, p. 633). In order to avoid being stuck into an infinite loop, because clearing the source peg to move disk i will block the target peg and vice versa, a stack of subgoals is maintained which allows the strategy to be rescued.

KR The strategy described in Karat (1982) which combines elements of domain-specific knowledge into a general problem-solving framework. The strategy adopts a limited look-ahead: if the movement of the LOOP disk from its source to the target peg is blocked by only the smallest two disks, the task of moving the small disks on the third peg is considered as trivial, and the moves are immediately executed.

AT In addition to implementing the above mentioned strategies, we utilized also the activation-based model of memory for goals (Altmann & Trafton, 2000)[4]. The model adopts the strategy of Anderson & Lebiere (1998), but stores goals as ordinary declarative memory elements instead of caching them in the architectural goal stack, and uses a strengthening process for encoding and priming from cues for retrieval.

As previously mentioned, all the strategies are sub-optimal, i.e. they do not generally reach the solution with the minimum number of moves, a performance that also our participants were seldom (i.e., 12% of the times in the accuracy, and 5% in the effort condition) able to make.

Procedure and Results

We executed a simulation of all the strategies on the TOH problems used in the experiment.

We decided to compare the strategies only on their capacity to predict the number of errors made by the participants. Additional assumptions and parameter tuning would be required to model also the times. Therefore, we preferred to stick to a very conservative simulation policy.

The trial-by-trial results of the simulation are presented in Table 1. The table shows the strategies that, in each trial, predicted a number of errors falling into the 99% confidence intervals (CI) computed from the experimental data.

Table 1: Trial-by-trial simulation results.

Group	Trial							
	1	2	3	4	5	6	7	8
Accuracy	KR SP AT	KR SP	SP AT	KR SP	SP	SP AT	KR SP	KR SP AT
Effort	S2 KR SP	KR SP	KR SP	KR	KR	KR		KR SP

The global fit of the three best strategies (SP, KR and AT)— measured using the mean absolute difference (MAD), the root mean square error (RMSE) and the percentage of trials in which the prediction of the model is within the 99% CI (P99CI)— is presented in Table 2.

Table 2: Simulation results for the best strategies.

Strategy	Group	MAD	RMSE	P99CI
KR	accuracy	3.098	3.615	62.5%
SP	accuracy	2.899	3.239	100%
AT	accuracy	5.093	5.763	50%
KR	effort	2.264	3.074	87.5%
SP	effort	5.992	7.111	50%
AT	effort	9.382	10.112	0%

The best fitting strategies are SP in the accuracy condition and KR in the effort condition. The AT

[4] We thank Erik Altmann for making the model available and allowing us to use it in the simulation study

strategy yields good results on half of the trials in the accuracy condition. The selective search strategies are notable to achieve a good fit: only the use of SS2 in the first trial of the effort condition cannot be excluded.

Discussion

The basic conclusion that can be drawn from the simulation is that the results are mainly in compliance with the contingent and adaptive selection hypotheses.

The perceptual strategy is actually more accurate but probably more effortful than the Karat's strategy (that does not require expensive recursive operations). The Altmann & Trafton's model is more accurate than the other two strategies, but probably more expensive than the Karat's model.

Further simulations, using model-tracing and time data, should provide additional supporting evidence.

Conclusions

A preliminary support has been gained for the contingent and adaptive nature of strategy selection in the TOH. On this basis, we suggest that it is important to pay attention to the problem solving factors affecting the accuracy vs. effort trade-off, due to their influence on the strategy selection.

Many other issues must be cleared to obtain a deeper understanding of the selection processes in the TOH and in similar well-structured problems. In this context, we regard as especially important the transition towards more detailed, cognitively grounded strategies to further constrain and specify the existing models, and to allow more detailed comparisons.

This process could yield both the redesign of old strategies and the definition of new ones. Altmann & Trafton (2000) offered a first important contribution with their memory-based model of the Anderson & Lebiere (1998) strategy. We think that a closer analysis and experimental investigation of the attentional and perceptual processes in the TOH could produce significative advances in our understanding of the cognitive processes underlying the solution strategies.

References

Ahlum-Heath, M. E. & DiVesta, F. J. (1986). The effect of conscious controlled verbalization of a cognitive strategy on transfer in problem solving. Memory & Cognition, 14, 281-285.

Altmann, E. M. & Trafton, J. G. (2000). An activation-based model of memory for goals. Manuscript submitted for publication.

Anderson, J. R. (1990). The adaptive character of thought. Hillsdale, NJ: Erlbaum.

Anderson, J. R., Kushmerick, N. & Lebiere, C. (1993). The tower of Hanoi and goal structures. In J. R. Anderson (Ed.), Rules of the mind. Hillsdale, NJ: Erlbaum.

Anderson, J. R. & Lebiere, C. (1998). The atomic components of thought. Hillsdale, NJ: Erlbaum.

Anzai, Y. & Simon, H. A. (1979). The theory of learning by doing. Psychological Review, 86, 124-140.

Christensen-Szalanski, J. J. J. (1998). Problem-solving strategies: a selection mechanism, some implications, and some data. In L. R. Beach (Ed.), Image Theory. Hillsdale, NJ: Erlbaum.

Davies, S. P. (2000). Memory and planning processes in solutions to well-structured problems. The Quarterly Journal of Experimental Psychology, 53A, 896-927.

Fum, D. & Del Missier, F. (2000). Adaptive spatial planning. Proceedings of the Seventh Annual ACT-R Workshop. Pittsburgh: Carnegie Mellon University.

Goel, V. & Grafman, J. (1995). Are the frontal lobes implicated in "planning" functions? Interpreting data from the tower of hanoi. Neuropsychologia, 33, 623-642.

Karat, J. (1982). A model of problem solving with incomplete constraint knowledge. Cognitive Psychology, 14, 538-559.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A. & Wager, T. D. (2000). The unity and diversity of executive functions and their contribution to complex "frontal lobe" tasks: a latent variable analyisis. Cognitive Psychology, 41, 49-100.

Payne, JW., Bettman JR., & Johnson, EJ. (1993). The adaptive decision maker. New York: Cambridge University Press.

Ruiz, D. & Newell, A. (1989). Tower-noticing triggers strategy-change in the Tower of Hanoi: a Soar model. Proceedings of the Eleventh Annual Conference of the Cognitive Science Society (pp. 522-529). Hillsdale, NJ: Erlbaum.

Simon, H. A. (1975). The functional equivalence of problem solving skills. Cognitive Psychology, 7, 268-288.

Stinessen, L. (1985). The influence of verbalization on problem solving. Scandinavian Journal of Psychology, 26, 342-347.

VanLehn, K. (1991). Rule acquisition events in the discovery of problem solving strategies. Cognitive Science, 15, 1-47.

Self-Organising Networks for Classification Learning from Normal and Aphasic Speech

Sheila Garfield, Mark Elshaw and Stefan Wermter
University of Sunderland
Informatics Centre, SCET
St. Peter's Way
Sunderland SR6 0DD, United Kingdom
Email: Stefan.Wermter@sunderland.ac.uk
Phone: +44 191 515 3279
Fax: +44 191 515 3553

Abstract

An understanding of language processing in humans is critical if realistic computerised systems are to be produced to perform various language operations. The examination of aphasia in individuals has provided a large amount of information on the organisation of language processing, with particular reference to the regions in the brain where processing occurs and the ability to regain language functionality despite damage to the brain. Given the importance played by aphasic studies an approach that can distinguish between aphasic forms was devised by using a Kohonen self-organising network to classify sentences from the CAP (Comparative Aphasia Project) Corpus. We demonstrate that the different distributions of words in aphasics types may lead to grammatical systems which inhabit different areas in self-organising maps.

Introduction

The examination of neural language processing is of importance as it offers the opportunity for producing realistic computerised language systems and a comprehension of the underlying biological mechanisms and constraints involved. One technique that has proved useful for identifying the organisational arrangement of language processing is the examination of aphasia. Aphasia is the inability to perform one or more cognitive language functions due to damage to the brain. The typical causes of aphasia are brain tumours, strokes, head injuries and infections. Although this is a rough simplification, the two most common types of aphasia are Broca's and Wernicke's aphasia.

Broca's Aphasia: Subjects with damage to the Broca's area of the cerebral cortex have problems creating spoken responses. These responses are often grammatically incorrect, effortful, laboured, come in bursts and have a restricted vocabulary. Furthermore, verbs are often missed out or replaced by the nominal form in spontaneous speech. However, many individuals with this condition can perform language processing functions such as language comprehension, dealing with non-reversible sentences, object and verb recognition and the identification of semantic and verb errors. Table 1 provides examples of typical spontaneous speech from Broca's aphasics [Wermter, Panchev and Houlsby (1999), Marshall, Pring and Chait (1998) and Brendt and Caramazza (1999)].

Wernicke's Aphasia: Although individuals with Wernicke's aphasia have problems understanding language and producing sentences that are meaningful, they can produce fluent phrases that have a reasonable syntactic, grammatical and symbolic structure [Chen and Bates (1998) and Wermter, Panchev and Houlsby (1999)]. Table 2 provides examples of spontaneous speech from Wernicke's aphasics.

An approach previously used to distinguish aphasic forms is recurrent neural networks [Wermter, Panchev and Houlsby (1999)]. Such networks can represent long term memory and context using recurrent connections and extracting the appropriate context from inputs. In the simple recurrent network outlined by Elman (1990) the context layer stores the activations of the hidden layer units for one time step and passes them back to the hidden layer units on the next step. Typically there is a one-to-one relationship between the number of units in the context layer and in the hidden layer [Spitzer (1999)]. This offers the opportunity to recycle information from multiple time steps and to identify temporal relationships. As the hidden layer receives inputs from both the input and the context layer, patterns should have an impact across time and context be learned.

However, there are certain drawbacks with recurrent neural networks which led us to consider an alternative approach. Recurrent neural networks are a *supervised* learning approach that do not perform in a manner that is close to neural networks in the human brain. Therefore in this paper we used *unsupervised* self-organising networks that can identify categories, features and regularities using unsupervised learning in a manner closer to the cerebral cortex. In this paper we analyse spoken language from Broca's aphasics, Wernicke's aphasics and normal patients. We demonstrate that the different distributions of words in aphasics types may lead to grammatical systems which inhabit different areas in self-organising maps.

Location of Aphasia and Language Function

The examination of aphasics provides some indication of how language processing is organised and the form that language recovery takes. A language processing model that has been established from studying the location of damage in the cerebral cortex of aphasics is that the human brain performs diverse language processing opera-

Table 1: Typical spontaneous speech from Broca's aphasics.

Normal phrase	Broca's aphasic response
A boy is giving the ball to the man	A boy is ... the ball
A monkey is eating a banana	Monkey ... banana
Chrysanthemum	Chrysa...mum...mum
Cat cries	Cat tears

Table 2: Typical spontaneous speech from Wernicke's aphasics.

Typical Wernicke's aphasic responses
They are running a swimming water and snow
The boy is running he is talking to the it is a cat
It is a cat and he is talking the flower

tions. According to Taylor (1999) and Dodel, Hermann and Geisel (1999) the cortex is made up of various somewhat overlapping regions which are responsible for cognitive language sub-operations. In order to produce the final language functions there is a need to coordinate and combine the outcomes of the appropriate regions. According to Reilly (2001) the brain performs as a group of collaborating specialists, none of which can deal with a difficulty alone, but only do so when each cooperates. In the brain it is possible to deal with a complex difficulty by splitting the task into smaller elements and coordinating these elements. The uniqueness of the human brain does not come from the number of neurons but the structural complexity. It has been identified that the module approach can offer re-usability by having a region of the brain doing the same processing activity as part of many different cognitive functions [Reilly (2001)].

In terms of the actual functions that are associated with diverse regions of the cerebral cortex a few examples will now be outlined. When Binder, Frost, Hammeke, Cox, Rao and Prieto (1997) required individuals to state whether an animal was native of America and used by humans, different principal regions of the cerebral cortex were established as responsible for the language processing involved: (i) an area incorporating the superior temporal sulcus, middle temporal gyrus and parts of the inferior temporal gyrus; (ii) sections of the inferior and superior frontal gyri, the middle frontal gyrus and the anterior cingulate; (iii) angular gyrus; and (iv) a region containing the posterior cingulate and gyrus zones.

Silent word generation starting with a certain letter takes place in Broca's and Wernicke's areas and sections of the left frontal, temporal and parietal lobes [Papke, Hellmann, Renger, Morgenroth, Knetcht, Schuierer and Petersen (1999)] and the resolution of whether two words belong to the same semantic group involves increased activity in the superior frontal gyrus and frontal gyrus [Shaywitz, Shaywitz, Pugh, Constable, Skudlarski, Fulbright, Bronen, Fletcher, Shankweiler, Katz, Gore (1995)]. Finally, the process of generating verbs out loud was found by Xiong, Rao, Gae, Woldroff, Fox (1998) and Raichle, Fiez, Videen, Macleod, Pardo, Fox, Petersen (1994) to be associated with areas of the left posterior temporal cortex, right anterior cingulate, inferior frontal gyrus, Broca's area, left superior temporal gyrus, cingulate gyrus, inferior temporal gyrus and the occipital gyri.

The examination of aphasia has assisted in creating models of the form that the recovery of language processing takes in the brain. Examinations of the brain following death have identified injuries to parts of the cerebral cortex in normally functioning individuals which should have produced aphasia. This led to the view from Karbe, Thiel, Weber-Luxenburger, Herholz and Heiss (1998), Basso, Gardelli, Grassi and Mariotti (1989) and Capp, Perani, Grassi, Bressi, Alberoni, Franceschi, Bettinardi, Todde, and Frazio (1997) that language functions are recovered through regeneration of the damaged tissue or the redistribution of functionality to other regions of the brain that are operationally linked but not required in healthy individuals.

There is mixed research evidence for the time it normally takes for repair of injured tissue. However, researchers have found than redistribution of functionality to new regions of the brain can take longer and repair of the left superior temporal gyrus occurs over numerous months following the injury [Mimura, Kato, Kato, Santo, Kojima, Naeser and Kashima (1998) and Weiller, Isensee, Rijntjes, Huber, Müller, Bier, Dutschka, Woods, Noth and Diener (1995)]. As early as in the 19th Century Gower determined that individuals who lost speech due to damage to the left hemisphere were able to recover it through interaction with the right hemisphere. The region of the right hemisphere analogous to Broca's area and the right perisylvian have taken over the functions associated with the Broca's and Wernicke's areas respectively when they are injured. According to Reggia, Shkuro and Shevtsova (2000) the reorganisation of the brain regions responsible for language explains the remarkable capacity to recover from injury and robust,

fault-tolerant processing. So in summary several brain regions may be involved with aphasia, even though at a highest level often a distinction of Broca's and Wernicke's aphasia has been made in the past.

Classification of Aphasia using Self-Organising Networks

As aphasia studies provide a significant amount of relevant information regarding the organisation of brain processing, there is a motivation to develop an approach to classify interviewed subjects to distinguish the aphasia form they have.

Method Overview

The language transcripts used for the training and test data sets for a self-organising network were obtained from the CAP Corpus [Bates, Fredrici and Wulfeck (1987a and 1987b)]. The CAP Corpus is made up of transcripts of English-speaking subjects that are divided into three groups: Broca's aphasia, Wernicke's aphasia and a control group of healthy people. The language transcripts were produced using a variation of the "given-new" picture description task of MacWhinney and Bates. In this task subjects were shown nine sets of three pictures and were asked to describe them (see Table 3). The transcripts contained the subject's response and the morphemic coding. We used the coding from a previous study by Wermter, Panchev and Houlsby (1999). This maps the morphemic coding of the corpus patterns using the following syntactic descriptors: DET (Determiner), N (Noun), N-PL (Plural), PRO (Pronoun), PREP (Preposition), ADJ (Adjective), CONJ (Conjunction), V (Verb), V-PROG (Progressive), AUX (Auxiliary Verb), ADV (Adverb), ADJ-N (Numeric).

Unsupervised Learning

The self-organising network that was used consists of an input and an output layer, with every input neuron linked to all the neurons in the output layer [Spitzer (1999), Hecht-Nielsen (1990), Kohonen (1997) and Anderson (1999)]. A self-organising network can be used by itself or as a layer of another neural network. Input data is presented one sample at a time and the nodes compete against each other. The Kohonen layer creates a topographical representation of the critical characteristics of the input by creating a pattern of active and inactive units (see Figure 1). The activation of the units are calculated by multiplying the input from each input unit by its related synaptic weight and summing all the inputs for a specific unit.

Learning in self-organising networks is performed by updating the links between the input layer and the output layer via a form of Hebbian learning. Self-organising networks attempt to depict the input data with a set of models, with similar words and concepts producing models that activate the units in the output layer that are close together.

Fitting of model sectors is performed by a sequential regression procedure, where $t = 1, 2, \ldots$ is the step index: For every sample $x(t)$, the winner index c is established by the condition

$$i, \;\; x(t) - m_c(t) \qquad x(t) - m_i(t)$$

Once this has occurred, every model vector m_i or a subgroup of them that belong to units centred around unit $c = c(x)$ are altered as

$$m_i(t \quad 1) = m_i(t) \quad h_{c(x),i}(x(t) - m_i(t))$$

The 'neighbourhood function' $h_{c(x),i}$ defines those units that are to be updated.

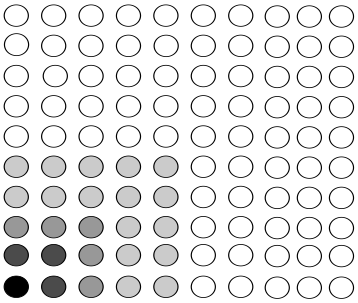


Figure 1: A representation of the activity maps of a self-organising network - The darker the neuron the greater the activation.

The self-organising network architecture considered to classify aphasic types contained 100 units (10 x 10) in the output layer. Using a different training/test set pair a self-organising network was trained and tested using the following approach. A network was trained over 1000 epochs using 80 phrases for each of the three aphasic types (Wernicke's aphasics, Broca's aphasics and a healthy control group) that were produced from the CAP Corpus. So in total there were 240 phrases. The location of each of these training phrases on the self-organising maps was identified based on the units that had the highest activation. The trained network was then tested by identifying where on the map 80 unseen phrases per aphasic type are positioned and the degree of symmetry between the training and test samples. The objective was to test if the phrases for Broca's and non-Broca's aphasics are located in different regions of the map and whether the network is able to generalise by placing the test phrases for the two groups in the same regions as the training ones. If the same unit has the highest activation level for phrases from both groups the unit is allocated to the aphasic type that has the most phrases associated with it. The grouping of Wernicke's aphasics with the healthy control group is motivated by the observation that Wernicke's aphasics often use correct syntax like the healthy control group while Broca's aphasics do not.

To remove any bias in classification and clustering the test/training phrases are based on the first six words of

Table 3: Picture series.

Syntactic Description	Sentences
DET N AUX V-PROG	A bear/mouse/bunny is crying.
DET N AUX V-PROG	A boy is running/swimming/skiing.
DET N AUX V-PROG DET N	A monkey/squirrel/bunny is eating a banana.
DET N AUX V-PROG DET N	A boy is kissing/hugging/kicking a dog.
DET N AUX V-PROG DET N	A girl is eating an apple/cookie/ice-cream.
DET N V PREP DET N	A dog is in/on/under a car.
DET N V PREP DET N	A cat is on a table/bed/chair.
DET N AUX V-PROG DET N PREP DET N	A lady is giving a present/truck/mouse to a girl.
DET N AUX V-PROG DET N PREP DET N	A cat is giving a flower to a boy/bunny/dog.

Table 4: Three word phrases for the aphasic types and their numeric representation.

Aphasic Type	Phrases	Syntactic Description	Numeric Represention
Broca's Aphasic	Banana three eat	NOUN ADJ-N VERB	1100 1010 0010
Broca's Aphasic	Boy is sport	NOUN AUX NOUN	1100 0100 1100
Wernicke's Aphasic	Little small here	ADJ ADJ PREP	0101 0101 1001
Wernicke's Aphasic	Squirrel with banana	NOUN PREP NOUN	1100 1001 1100
Healthy Control	The banana eating	DET NOUN V-PROG	0110 1100 1000
Healthy Control	A young boy	DET ADJ NOUN	0110 0101 1100

the sentences. A sliding window of three words that moves along one word at a time is used to create the final training/test three word phrases. Hence, if a transcript includes a sentence *"The monkey is sitting down eating a small yellow banana."* the first six words obtained are *"the monkey is sitting down eating"* and two of the training/test phrases are *"the monkey is"* and *"monkey is sitting"*. Since every word of these phrases is represented by a four digit binary number, the input layer for the network architecture has twelve units. The binary representations for the word are Determiner (0110), Noun (1100), Plural (0011), Pronoun (0111), Preposition (1001), Adjective (0101), Conjunction (1011), Verb (0010), Progressive (1000), Auxiliary Verb (0100), Adverb (0001) and Numeric (1010). Table 4 shows typical reponses of the aphasic types and the numeric representations that were input for the self-organising networks.

Results

Figures 2 and 3 show that it is possible to identify clear regions of the self-organising networks that are associated with the Broca's aphasic test phrases for both the test and training data. For Broca's aphasics there are two clear regions of the map, which is an indication that two forms of the condition might exist. For the two maps the Wernicke's aphasic/healthy control group are distributed around the rest of the map. When considering the test and training sample locations it is clear that the areas of the map associated with the test Broca's aphasics are very similar to the training ones. In many cases the cells with the highest activiation are exactly the same for the training and test samples. Therefore, unsupervised self-organising networks are a suitable alternative to supervised approaches for classifying aphasic types.

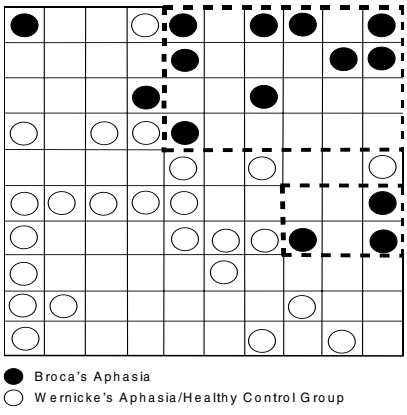


Figure 2: The regions on the self-organising map for a network based on 12 input and 100 output layer units associated with the second training set phrases for the aphasic types.

It is often the case when neural networks are trained to learn grammatical structures that two classes of examples are used; grammatically correct and incorrect phrases. The self-organising network architecture used in this paper is more general than these networks as it can identify three grammatical phrase structures, where the

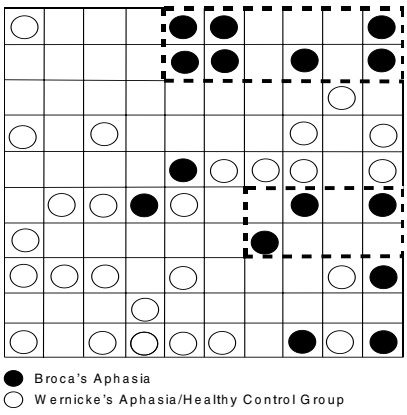


Figure 3: The regions on the self-organising map for a network based on 12 input and 100 output layer units associated with the second test set phrases for the aphasic types.

test phrases contain both typical and non-typical grammatical structures. Since phrases for the healthy control group/Wernicke's aphasics and Broca's aphasics are located at different regions on the self-organising maps it may be possible to develop a model of how the brain represents and processes grammatical structures of different individual types [Zurif, Swinny, Parther, Solomon, and Bushell (1993), Hartsuiker and Kolk (1998) and Marshall, Pring and Chiat (1998)].

The results in our experiments indicate that unsupervised networks are a suitable alternative to supervised approaches for classifying aphasic types. In terms of cognitive science the results show that while the spoken output of Broca's aphasics has a very distinct grammatical structure, healthy individuals and Wernicke's aphasics are much closer. This supports the view that language production may be based on a modular but interactive approach associated with particular regions of the brain and that correct grammatical construction is dependent on Broca's area.

By identifying two clear zones of the output maps associated with Broca's aphasics these could be associated with different degrees of injury and performance. If this is the case the different individuals in the two groups could provide the basis of a computational model of different levels of Broca's injury and hence of recovery. A final issue for consideration is why those classified as Broca's aphasic by the self-organising network failed to recover functionality by either tissue recovery or functional redistribution. A case examination of these individuals might provide information on the factors that are significant in functional recovery such as age, extent of injury and the type and level of medical intervention following injury.

The approach in this paper for classifying different aphasic types using a self-organising network was based on the difference between the grammatical constructs produced. This is an important step in our research with our overall aim being to incorporate other spoken language characteristics such as semantics and vocabulary level into the classification process by using a set of self-organising nets. The impact of that would be to produce a benchmark approach to classify many more aphasic types using a self-organisation approach and so provide cognitive scientists with a powerful diagnostic tool.

An additional advantage to cognitive scientists from the extented classifier is the removal of the subjective manner by which researchers include and exclude aphasics from pooled studies. For example, when considering if Broca's aphasics can deal with reversable sentences Brendt and Caramazza (1999) state that the percentage that cannot deal with these sentences is much less than those identified by Grodzinsky, Piñango, Zurif and Drai (1999) from the examination of the same pooled studies. Brendt and Caramazza (1999) add that the difference comes from Grodzinsky, Piñango, Zurif and Drai (1999) willingness to exclude Broca's aphasics. It is argued that they are not true Broca's aphasics. Finally, this system should offer an indication of the underlying organisational properties of language in the brain and so assist with the development of computational hybrid neural processing models [Wermter and Sun (2000) and Wermter and Meurer (1997)].

Conclusion

Studying individuals that have aphasia has provided a great deal of information connected with the nature of language processing and how the brain is able to recover language functionality following injury. By using self-organising network architectures it is possible to distinguish between a control group of healthy individuals/Wernicke's aphasics and Broca's aphasics using sentences from the CAP Corpus. One possible reason for the self-organising network's ability to separate the inputs into these two groups is their capacity to learn the grammatical structure produced by these aphasic types, which typically for Broca's aphasics are grammatically incorrect and for Wernicke's aphasics/healthy individuals are grammatically correct.

Acknowledgments

This research has been funded in part by EPSRC GR/M56555/01 awarded to Prof. Wermter.

References

Anderson, B. Kohonen neural networks and language. *Brain and Language*, 70:86–94, 1999.

Basso, A., Gardelli, M., Grassi, M. & Mariotti, M. The role of the right hemisphere in recovery from aphasia: Two case studies. *Cortex*, 25:555–566, 1989.

Bates, E., Friederici, A. & Wulfeck, B. Grammatical morphology in aphasia: Evidence from three languages. *Cortex* 23:545–574, 1987a.

Bates, E., Friederici, A. & Wulfeck, B. Sentence comprehension in aphasia: A cross-linguistic study. *Brain and Language*, 32:19–67, 1987b.

Binder, J., & Frost, J., Hammeke, T., Cox, R., Rao, S. & Prieto, T. Human brain language areas identified by functional magnetic resonance images. *The Journal of Neuroscience*, 17(1):280–288, 1997.

Brendt, R. & Caramazza, A. How regular is sentence comprehension in broca's aphasia? it depends on how you select the patients. *Brain and Language*, 64(2):231–256, 1999.

Capp, S., Perani, D., Grassi, F., Bressi, S., Alberoni, M., Franceschi, M., Bettinardi, V., Todde, S. & Frazio, F. A PET Follow-up Study of Recovery after Stroke in Acute Aphasics. *Brain and Language*, 56:55–67, 1997.

Chen, S. & Bates, E. The dissociation between nouns and verbs in Broca's and Wernicke's aphasia: Findings from Chinese. *Aphasiology*, 12(1):5–36, 1998.

Dodel, S., Hermann, J. & Geisel, T. Stimulus-independent data analysis of fMRI data. In Wermter, S., Austin, J. & Willshaw, D., editors, *EmerNet: International Workshop on Emergent Neural Computational Architectures Based on Neuroscience*, pages 7–10. EmerNet, 1999.

Elman J. Finding structures in time. *Cognitive Science*, 14:179–211, 1990.

Grodzinsky, Y., Piñango, M., Zurif, E. and Drai, D. The Critical Role of Group Studies in Neuropsychology: Comprehension Regularities in Broca's Aphasia, *Brain and Language*, 67:134–147, 1999.

Hartsuiker, R. & Kolk, H. Syntactic Facilitation in agrammatic sentences production. *Brain and Language*, 62:221–254, 1998.

Hecht-Nielsen, R. *Neurocomputing*. Addison-Wesley, Reading MA, 1990.

Karbe, H., Thiel, A., Weber-Luxenburger, G., Herholz, K. & Heiss, W. Brain plasticity in poststroke aphasia: What is the contribution of the right hemisphere? *Brain and Language*, 64:215–230, 1998.

Kohonen, T. *Self-Organizing Maps*. Springer Verlag, Heidelberg, 1997.

Marshall, J., Pring, T. & Chait, S. Verb Retrieval and Sentence Production in Aphasia. *Brain and Language*, 63(2):159–183, 1998.

Mimura, M., Kato, M., Kato, M., Santo, Y., Kojima, T. and Naeser, M. and Kashima, T. Prospective and Retrospective Studies of Recovery in Aphasia: Changes in Cerebral Blood Flow and Language Functions. *Brain*, 121: 2083–2094, 1998.

Papke, K., Hellmann, T., Renger, B., Morgenroth, C., Knecht, S., Schuierer, G. & Reimer, P. Clinical applications of functional MRI at 1.0 t: motor and language studies in healthy subjects and patients. *European Radiology*, 9(2):211–220, 1999.

Raichle, M., Fiez, J., Videen, T., MacLeod, A., Pardo, J., Fox, P. & Petersen, S. Practice-related changes in human brain functional-anatomy during nonmotor learning. *Cerebral Cortex*, 4(1):34–54, 1994.

Reggia, J., Shkuro, Y. & Shevtsova, N. Emergent Specialization in Cerebral Regional Modules. In Wermter, S., Austin, J. & Willshaw, D., editors, *Proceedings of the Third International Workshop on Computational Architectures Intergrating Neural Networks and Neuroscience*, pages 11–14. EmerNet, 2000.

Reilly, R. Collaborative cell assemblies: Building blocks of cortical computation. In Wermter, S., Austin, J. & Willshaw, D., editors, In *Emergent Neural Computational Architectures based on Neuroscience*, Springer-Verlag, Heidelberg, Germany, 2001.

Shaywitz, B., Shaywitz, S., Pugh, K., Constable, R., Skudlarski, P., Fulbright, R., Bronen, R., Fletcher, J., Shankweiler, D., Katz, L. & Gore, J. Sex differences in the functional organisation of the brain for language. *Nature*, 373(6515):607–609, 1995.

Spitzer, M. *The Mind Within the Net: Models of Learning, Thinking and Acting*. MIT Press, Cambridge, MA, 1999.

Taylor, J. Images of the mind: brain images and neural networks. In Wermter, S., Austin, J. & Willshaw, D., editors, *Proceedings of the International Workshop on Emergent Neural Computational Architectures based Neuroscience*, pages 1–6. EmerNet, 1999.

Weiller, C., Isensee, C., Rijnties, M., Huber, W., Müller S., Bier, D., Dutschka, K., Woods, P., Noth, J. & Diener, C. Recovery from Wernicke's aphasia: A positron emissions tomographic study. *American Neurological Association*, 37(6):723–732, 1995.

Wermter, S., Panchev, C. & Houlsby, J. Language disorders in the brain: Distinguishing aphasia forms with recurrent networks. In *AAAI99 Conference Workshop on Neuroscience and Neural Computation*, pages 93–98, 1999.

Wermter S. & Meurer M. Building Lexical Representations Dynamically Using Artificial Neural Networks. *Proceedings of the International Conference of the Cognitive Science Society*, pages 802–807, Stanford, 1997.

Wermter, S. and Sun, R. *Hybrid Neural Systems* Springer-Verlag, Heidelberg, 2000.

Xiong, J., Rao, S., Gao, J., Woldorff, M. & Fox, P. Evaluation of hemispheric dominance for language using functional MRI: A comparison with positron emission tomography. *Human Brain Mapping*, 6:367–389, 1998.

Zurif, E., Swinney, D., Prather, P., Solomon, J. & Bushell, C. An on-line analysis of syntatic processing in Broca's and Wernicke's aphasia. *Brain and Language*, 45:448-464, 1993.

Rational imitation of goal-directed actions in 14-month-olds

György Gergely (113524.2623@compuserve.com)
Institute for Psychology of the Hungarian Academy of Sciences
1132 Budapest, Victor Hugo u. 18-22, Hungary

Harold Bekkering (H.Bekkering@ppsw.rug.nl)
Max Planck Institute for Psychological Research
Amalienstr. 33, D 80799 Muenchen, Germany

Ildikó Király (kiralyi@mtapi.hu)
Institute for Psychology of the Hungarian Academy of Sciences
1132 Budapest, Victor Hugo u. 18-22, Hungary

Abstract

The study sheds new light on the nature of imitative learning in 14-month-olds. It is demonstrated that while infants of this age can indeed imitate a novel means modelled to them, they do so only if the action is seen by them as the most rational alternative to the goal available within the constraints of the situation. The findings support the 'rational imitation' account over current 'imitative learning' or 'emulative learning' accounts in explaining re-enactment of goal-directed action in 14-month-olds.

Introduction

In a well-known study Meltzoff (1988, 1999) demonstrated that 14-month-olds are already capable of delayed imitation of a novel goal-directed action. Infants observed a salient novel action performed by an adult model on a black box with a translucent orange plastic panel for a top surface. The box had a light bulb hidden in it. The model leaned forward from the waist and touched the panel with his/her forehead as a result of which the box was illuminated. The infants were given the box only on a separate visit a week later when 67% of them imitated the salient novel action: they leaned forward themselves to touch the box with their forehead (see Figure 1); an action they would not spontaneously perform (as shown by a control base-line condition). This demonstration indicates the remarkably early presence of imitative learning. Meltzoff argues that 14-month-olds differentiate between the actor's goal (the visible outcome of the box lighting up) and the specific means (head-on-box) performed and "they imitate the means used, not solely the general ends achieved" (1995, p. 509). The present study addresses two important questions that arise in relation to Meltzoff's intriguing demonstration: 1) Why do infants imitate the specific novel action modelled? 2) Why don't they simply push the panel with their hand to achieve the outcome (this being a simpler, more familiar, and easier-to-perform action alternative available to them)?

In his work on the social transmission of tool use in chimpanzees, Tomasello (1999) differentiated between 'imitative learning' – which seems to be a human-specific capacity – and 'emulation learning' that is characteristic of nonhuman primates.

Figure 1: Touching the box with the forehead

Briefly, when primates observe a novel instrumental action that brings about an interesting outcome, they seem to focus on the salient outcome only without differentiating it from the particular means used. This is suggested by the fact that when they attempt to bring about the same outcome themselves – in contrast to young children – they do not directly imitate the specific means modelled. Rather, they perform a series of motor actions directed to the outcome that are already available in their motor repertoire, until – through a process of trial-and-error learning - they hit upon the same effective means that was modelled for them, as if 'reinventing' it by chance.

Tomasello (1999) points out that if infants used emulation learning in the Meltzoff situation, one could expect that instead of imitating the novel and unfamiliar

'head-on-box' action, they would tend to perform a simpler, more natural, and already familiar motor action to achieve the outcome: they would touch the box with their hand (but, apparently, they did not). Therefore, Tomasello (1999) argues that infants in the Meltzoff study "understood a) that the adult had the goal of illuminating the light; b) that he chose one means for doing so, from among other possible means; and c) that if they had the same goal they could choose the same means – an act in which the child imagines herself in the place of the other" According to this simulationist account "imitative learning of this type thus relies fundamentally on infants' tendency to identify with adults…" (p. 82).

At first sight, infants' readiness to faithfully imitate the novel and unfamiliar 'head-on-box' action also seems unexpected in the light of Gergely and Csibra's recent theory of the one-year-old's 'naïve theory of rational action' (Gergely, Nádasdy, Csibra, & Bíró, 1995; Gergely & Csibra, 1997; Csibra & Gergely, 1998). In a series of habituation studies these authors and their colleagues (Gergely et al., 1995; Csibra, Gergely, Bíró, Koós, & Brockbank, 1999) demonstrated that 9 to 12-month-old infants (but not 6-month-olds) can already interpret the behaviour of an abstract computer-animated figure as a goal-directed rational action. For example, infants were habituated to a visual event in which a small circle repeatedly approached and contacted a larger circle by jumping over a rectangular figure (the 'obstacle') that was placed in between them. During the test phase, the 'obstacle' was removed, and infants were presented with either of two events. In the 'old action' event (non-rational approach) the small circle performed the same jumping approach as before to get to the large circle, even though - for adult observers – this jumping-over-nothing action did not seem a 'sensible' goal approach given the absence of the 'obstacle'. In the 'new action' event (rational approach) the small circle performed a novel but (for adults) 'sensible' action: it approached the large circle by following the most direct horizontal straight-line pathway that has become available leading to the large circle. Corresponding to adult intuitions, 9- and 12-month-olds looked longer at the non-rational 'old action' event than at the (rational) 'new action' event, while showing no dishabituation to the latter.

According to Gergely and Csibra's theory this finding demonstrates that when interpreting a goal-directed behaviour, one-year-olds evaluate the rationality of the particular action as a function of the visible goal and the physical constraints of the actor's situation (here the presence of the 'obstacle'). When the situational constraints change (i. e., when the 'obstacle' is removed), infants can infer what particular novel action the actor ought to perform in the new situation to achieve the goal in the most rational or efficient manner. It is hypothesised that in doing so infants rely on the inferential principle of rational action that assumes that to achieve its goal an agent will choose to perform the most rational action available given the constraints of the situation (Gergely & Csibra, 1997; Csibra & Gergely, 1998).

Extending this theory to imitative learning situations one would expect infants to imitate the model's novel means only if it appeared to them to be the most rational or efficient alternative to the goal within the constraints of reality. On this assumption, however, it is not immediately clear why Meltzoff's subjects would consider the novel 'head-on-box' action as the most rational means to the goal, when clearly there is a much simpler, more familiar and for them obviously easier-to-perform motor alternative: they could touch the box simply by placing their hands on it ('hand-on-box' action). Why do they imitate the novel 'head-on-box' action then?

To solve this riddle, we hypothesised that it is possible that the action modelled by Meltzoff contained certain situational features that allowed infants to 'rationalize' the 'head-on-box' action as the most efficient alternative available to the goal. In particular, it seems possible that infants noticed and interpreted the fact that while the model's hands were free to act, s/he nevertheless chose to touch the box with his/her forehead rather than with his/her hands. Assuming that the adult is a 'rational agent', the infants may have concluded that 'there must be a good reason' for this choice, and that the 'head-on-box' action must have advantages over the simpler-looking 'hand-on-box' action in achieving the goal. Therefore, when getting a chance to reproduce the effect, the infants themselves would opt to perform the novel 'head-on-box' action that had been inferred to be the most rational alternative to the goal.

What would happen if the model's hands were visibly occupied while s/he was performing the 'head-on-box' action? This would make it explicit that in the given situation the simpler 'hand-on-box' action is not available to the model, and so the performed 'head-on-box' action would clearly appear to be the most rational alternative to the goal. What would infants do in this case, if after having observed the modelled 'head-on-box' action, we made the box available for them to act on? Note that here the situational constraints on available means would be different in the infant's case than in the case of the adult model, since, unlike the adult's, the hands of the infants would remain free to act. Therefore, while the modelled 'head-on-box' action may have seemed rational for the adult to perform, in case of the infants it would cease to be the most rational alternative available. For them there would clearly be a simpler and more rational means accessible in the form of the familiar and well-practiced 'hand-on-box' action.

Therefore, on the basis of the 'principle of rational action' we would expect that in this situation infants would not faithfully imitate the adult's 'head-on-box' action, but rather they would be more likely to touch the box with their hands: an action that is more rational given the constraints of their own situation.

In sum: our 'rational imitation' account outlined above differs from Meltzoff's and Tomasello's 'imitative learning' accounts in two significant respects. First, the 'imitative learning' model, as it stands, predicts that infants would imitate the particular means modelled by an adult irrespective of whether the specific action is seen as the most rational alternative to the goal or not (cf. Nagell, Olguin, & Tomasello, 1993). In contrast, our 'rational imitation' account emphasizes that infants do not imitate faithfully or automatically an adult model's goal-directed action. Rather, they would first evaluate the modelled behaviour from the point of view of the 'principle of rational action' and imitate it only if they managed to 'rationalize' it as the most efficient alternative to the goal available given the constraints of the particular situation. Second, the 'rational imitation' model predicts that infants will imitate the model's means that was judged to be rational only if the situational constraints of the adult model are similar to those of the infants. If the situational constraints are different, however, and there is a more rational alternative available to the infant that was not available to the model, infants are expected to perform this more rational means rather than imitating faithfully and automatically the specific action modelled by the adult.

We have modified the original Meltzoff (1988) situation in such a way that would allow us to test the above predictions.

Method

Subjects

We tested 30 14-month-old infants (+/- 1 week) in two experimental conditions. Three babies were dropped because they were not brought back for the second test, so overall we report data from 27 infants.

Procedure

The infants were brought to our lab twice with a one-week delay in between. On the first visit infants were seated in their mother's lap in front of a table that had 3 toy objects covered with cloths. (Here we are reporting data only for the 'magic box' object.) The experimenter sat at the other side of the table, while the infants were seated about one meter away from the table so that they could not reach the toys. The sessions were video taped from behind a one-way mirror. On the first visit the experimenter modelled the target act three times making sure that the infant paid attention.

The 'Hands free' condition (n=13) was a slightly modified[1] version of Meltzoff's (1988) original study. In this condition, even though the model's hands were visibly free, she did not use them. Instead, by leaning forward from the waist she touched the lamp on the box with her forehead ('head-on-box' action) and the lamp lit up. Note that in this situation the actor's reason for not using his free hands to touch the box is not directly demonstrated: it is only implied by her choice to use her head rather than her hand to light up the lamp.

In the 'Hands occupied' condition (n=14), before presenting the 'head-on-lamp' action the model, pretending to be freezing, told another experimenter that "she was cold and would like to have her blanket". After it was handed over to her, she wrapped it around her shoulders and held it tightly with both hands. (In the 'Hands free' condition the model also asked for her blanket, but then she put the blanket around her shoulders leaving her hands visibly free in front of her.) Note that in this condition the relevant situational constraints are different in the case of the model than in the case of the infant: while the hands of the adult were occupied, the hands of the infant were free. In both conditions the model went on to perform the very same 'head-on-box' action lighting the lamp by touching it with her forehead. (She repeated this three times.)

The test phase: Infants returned a week later. Sitting in their mother's lap they were allowed to act on the 'magic lamp' themselves. The model sat on the other side of the table as before. The infants actions were videotaped from behind a one-way mirror.

Data analysis and scoring

The video records of the test phase were scored by two independent observers who were uninformed as to which of the two conditions the subject belonged to. If the infant attempted to imitate the 'head-on-box' target action within a 20 sec time window s/he received 1 point, if s/he did not, s/he got 0 point. An attempt was defined as either touching the lamp with the head, or leaning forward in such a way that the subject's head approached the lamp within 10 cm or less (this is

[1] Our 'magic box' was slightly different from the one used by Meltzoff (1988) in that we have mounted a circular translucent table lamp on top of the box that could be activated by touch (see Figure 1). We have used this arrangement because in a pilot study identical to Meltzoff's experiment we noticed that the head and hair of the adult model often blocked the light effect from the infants when she touched the surface of the box with her forehead leaning over it. As a result some of the infants seemed to notice the light effect (showing surprise) only when they themselves touched the box during their second visit. By mounting the circular touch-sensitive table lamp on the box the resulting light effect was clearly visible to all infants already during the modeled action.

identical to Meltzoff's (1988) original criterion). The observers also coded the number of 'hand-on-box' actions and the number of times infants pointed to the model within the 20 sec time window. There was a 97% agreement between the two independent coders.

Results

To test our hypothesis that the different situational constraints on action in the two conditions influence the likelihood of the target action being imitated, we first compared the relative amount of imitated 'head-on-box' target acts in the two conditions. As Figure 2 shows, the two conditions differed significantly in this respect (Chi-square = 6.238 (df=1) p<.013). In the 'Hands free' condition 75% of the infants imitated the modelled 'head-on-lamp' action replicating Meltzoff's original result (he found 67% imitation). In contrast, when the model's hands were occupied ('Hands occupied' condition), only 27% of infants imitated the target act. The rest of the infants tried to light the lamp by touching it with their hands only.

Furthermore, there was a clear indication that the majority of infants who did not imitate the target act in the 'Hands occupied' condition did not fail to do so because they forgot the target act after the one week delay.

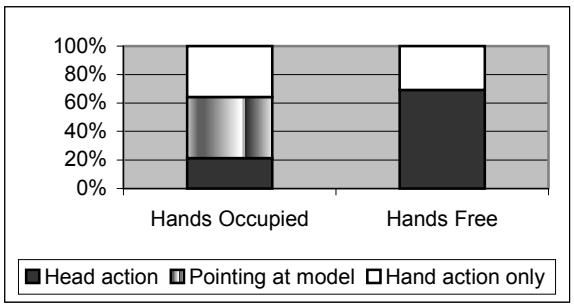


Figure 2:The amount of head and other types of actions in the two conditions

This is shown by the fact that 6 out of the 11 subjects (55%) not imitating the target act in this group produced a playful pointing gesture, pointing to the model (often smiling or giggling) (Figure 3a). This clearly indicates that they did recall the salient 'head-on-box' action of the model. In spite of this, however, they chose not to imitate, but proceeded to make the lamp light up by touching it with their hand (Figure 3b): a means that was simpler and more rational alternativa in their situation than the novel target act modelled. Furthermore, we found that all subjects in both conditions did produce at least once the 'hand-on-box' action within the 20 sec time window. In fact, the 'hand-on-box' action was typically performed more than once (Mean=2.1) by most subjects.

Figure 3a: Pointing to the model

Figure 3b: Touching with hand

Moreover, the large majority of infants (9 out of 12) who re-enacted the modelled action, performed the 'hand-on-box' action before imitating the 'head-on-box' action (Figure 4). Finally, in all cases where a 'hand-on-box' action was performed before the 'head-on-box' action, the hand-on-box' action was successful in bringing about the goal (i.e. the light was illuminated).

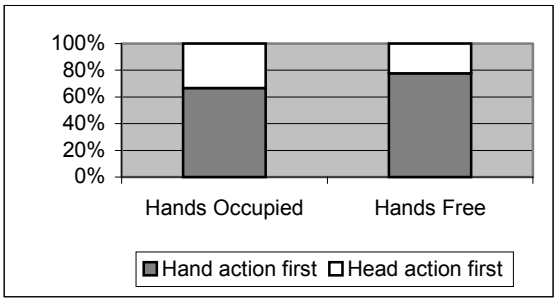


Figure 4: Relative order of hand action vs. head action among imitators

Discussion

The results provide support for the assumptions of the 'rational imitation' account. The differential degree of imitating the same target act found in the two conditions demonstrates that novel goal-directed actions modelled by an adult are not automatically imitated by 14-month-olds. The likelihood of reenacting a novel means observed was clearly a function of the infants'

interpretation of the rationality of the instrumental act in relation to the situational constraints on the actor's possible actions. We found that infants only imitated the 'head-on-box' action, if the contextual constraints of the adult's situation were the same as those of the infants themselves ('Hands free' condition). In this case 75% of the infants imitated the novel action, replicating Meltzoff's (1988) original finding. In contrast, when the model's hands were occupied ('Hands occupied' condition), the very same 'head-on-box' target act was imitated only by 27% of the infants. Given that their own hands were free to act, 73% of the 14-month-olds chose not to imitate the model in this condition, but performed a more rational alternative action available to them: they simply touched the lamp with their hand (Figure 3b).

We find especially informative the fact that in the 'Hands occupied' condition, in which the majority (73%) of the infants did not imitate the novel 'head-on-box' action, more than half of the non-imitating subjects pointed at the model while showing amusement (Figure 3a). This pointing gesture clearly indicates that after a week delay these infants successfully recalled the modelled 'head-on-box' action. Nevertheless, they chose to perform the (non-modelled) 'hand-on-box' action that was seen as a more rational means available to them in their own situation.

But why did the majority of infants reenact the novel 'head-on-box- action after having succeeded in lighting up the lamp by simply pushing it with their hands (Figure 4)? One possibility is that this imitative act served a communicative function: maybe to remind the model after a week that they remembered his funny action or to make him repeat his act. Alternatively, the reenactment may have served an epistemic function. Our data suggest that the infants inferred from seeing the model's free hands that 'there must be some reason' behind his choice to use his head instead of his hands to touch the box. Therefore, they may have expected the 'head-on-box' action to be in some (yet unknown) ways more advantageous. So maybe they reenacted the novel head action to discover the 'reason' behind the model's choice by experiencing the potential differences between the two alternative means. (We are currently running studies to test these hypotheses.)

To conclude:

1. The results successfully extend our 'naive theory of rational action' from the domain of action interpretation to the domain of action production and imitative learning.

2. We have demonstrated that evaluating the rationality of intentional action in relation to visible goals and situational constraints takes place at two different levels: a) during interpreting goal-directed actions performed by others (*ENCODING)*, and b) during selecting an appropriate motor response to achieve the same goal by the self (*RESPONSE GENERATION).*

3. Our findings show that re-enactment of a modelled goal-directed action is not an automatic process triggered by identification with a human actor. While identification may be involved in imitation, it is not sufficient to account for the differential pattern of re-enactment in our two conditions.

4. Instead, re-enactment of intentional action is a selective interpretative process driven by the inferential principle of rational action. Re-enactment takes place only a) if the action is judged as rational given the situational constraints of the *model*, and b) if the action is judged as potentially more rational than other available alternatives given the situational constraints of the *infant herself***.**

Acknowledgements

This research was carried out as part of an ongoing cooperation between the Institute for Psychology of the Hungarian Academy of Sciences and the Max Planck Institute for Psychological Research in Munich.

References

Csibra, G., & Gergely, G. (1998). The teleological origins of mentalistic action explanations: A developmental hypothesis. *Developmental Science*, *1*, 255-259.

Csibra, G., Gergely, G, Bíró, S., Koós, O., & Brockbank, M. (1999). Goal attribution without agency cues: the perception of 'pure reason' in infancy. *Cognition, 72*, 237-267.

Gergely, G., & Csibra, G. (1997). Teleological reasoning in infancy: The infant's naive theory of rational action. A reply to Premack and Premack. *Cognition, 63*, 227-233.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56*, 165-193.

Meltzoff, A. N. (1988). Infant imitation after a 1-week-delay: Long term memory for novel acts and multiple stimuli. *Developmental Psychology, 24,* 470-476.

Meltzoff, A. N. (1995). What infant memory tells us About infantile amnesia: Long term recall and deferred Imitation. *Journal of Experimental Child Psychology, 59,* 497-515.

Nagell, K., Olguin, K., & Tomasello, M. (1993). Processes of social learning in the tool use of chimpenzees (Pan troglodytes) and human children (Homo sapiens). Journal of Comparative Psychology, 107, 174-186.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press

The Right Tool for the Job: Information-Processing Analysis in Categorization

Kevin A. Gluck (kevin.gluck@williams.af.mil)
Air Force Research Laboratory, 6030 S. Kent St., Mesa, AZ 85212 USA

James J. Staszewski (jjs@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Howard Richman (howard@pahomeschoolers.com)
Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Herbert A. Simon (hs18@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Polly Delahanty (pd2w@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Abstract

Smith and Minda (2000) showed that mathematical approximations of several popular categorization theories could be fit equally well to the average "percentage of 'A' responses" in their meta-analysis of studies that used the 5-4 category structure. They conclude that the 5-4 category structure is not a useful paradigm for explaining categorization in terms of cognitive processes. We disagree with their conclusion, and contend instead that the problem lies with the data collection and analysis methods typically used to study categorization (in this and other category structures). To support this claim, we describe a recently completed study in which we collected and used a variety of converging data to reveal the details of participants' cognitive processes in a 5-4 category structure task.

The Smith and Minda (2000) Meta-Analysis

Recently, Smith and Minda (2000) reanalyzed 29 data sets (each set a particular condition in an experiment) collected from the experimental literature on categorization that used the 5-4 category structure.[1] Eight of the sets employed the stimuli called Brunswik faces. Others used yearbook photos (4 sets), geometric shapes (11 sets), verbal descriptions (3 sets), and rocketships (3 sets).

As shown in Table 1, each stimulus in this category structure has 4 binary features, whose combination creates 16 (2^4) different stimuli. The 5-4 structure splits this set into two linearly-separable groups.

In the acquisition phase of a typical category learning study, participants first learn to classify 9 of the 16 stimuli, 5 as A and 4 as B, as shown in the table. Each trial presents the 9 learning items, one at a time, in a random sequence, and the order changes from trial to trial. Participants classify each as "A" or "B" and the correct assignment for each stimulus is given as feedback. Typically, learning proceeds until a participant classifies all 9 stimuli correctly in a single trial. In the transfer test that follows, all 16 stimuli are presented to the participants and they classify each, now without feedback.

Table 1: The 5-4 category structure.

Stimulus (M&S, 1981)	Stimulus (S&M, 2000)	Feature F1	F2	F3	F4
Category A					
4A	A1	1	1	1	0
7A	A2	1	0	1	0
15A	A3	1	0	1	1
13A	A4	1	1	0	1
5A	A5	0	1	1	1
Category B					
12B	B6	1	1	0	0
2B	B7	0	1	1	0
14B	B8	0	0	0	1
10B	B9	0	0	0	0
Transfer					
1A	T10	1	0	0	1
3B	T11	1	0	0	0
6A	T12	1	1	1	1
8B	T13	0	0	1	0
9A	T14	0	1	0	1
11A	T15	0	0	1	1
16B	T16	0	1	0	0

Note. M&S = Medin & Smith; S&M = Smith & Minda. The feature structure for Medin & Smith's stimulus 4 is identical to that in Smith & Minda's stimulus A1, and so on.

Fitting the Data

For purposes of their meta-analysis, Smith and Minda (2000) used data from the transfer trial in each of these 29

[1] They speak of 30 sets, but two of their sets are obviously a duplication from an experiment by Medin and Smith (1981), for all 16 data points are identical for both sets.

data sets. Specifically, for each of the 16 stimuli, they computed the percentage of participants in each study who classified a stimulus as an 'A' stimulus. They then averaged these percentages over all 29 data sets to provide a global average, containing 16 data points, one for each stimulus. Thus, each data point represents the average percentage of participants (in those 29 studies) who categorized a particular stimulus as an 'A' stimulus. These data are displayed in Figure 1.

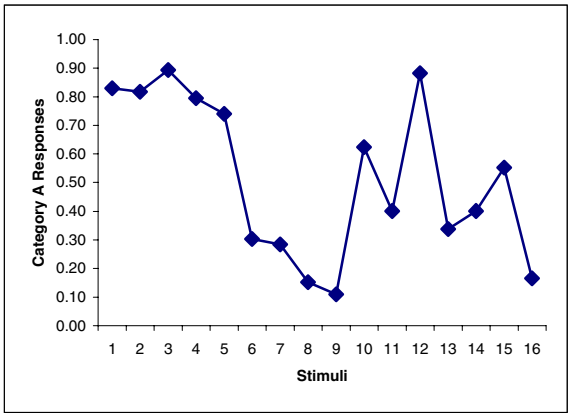


Figure 1. Average percentage of category 'A' responses from the Smith and Minda (2000) meta-analysis.

Smith and Minda (2000) next built a set of eight mathematical models, one for each of the theories they evaluated, and fitted each to the data for the 16 stimuli, adjusting the available parameters (4-6 free parameters per model) separately for best fit to each set of data. Then they averaged the set of 29 predicted performance profiles, to provide a set of 16 data points for comparison with the aggregated average of the actual data over all 29 studies. As Smith and Minda give a very clear and complete account of the functions they used and the fitting procedure, we need not repeat that information here. Table 2, adapted from Smith and Minda's Table 2, summarizes the fits to the data and the number of free parameters available to each model.

Table 2: Measures of Fit for Mathematical Models in Smith and Minda (2000)

Model	AAD	PVA	FP's
Additive prototype	0.091	0.838	4
Multiplicative prototype	0.069	0.890	5
Context	0.047	0.941	5
Additive exemplar	0.144	0.664	4
Fixed low sensitivity	0.149	0.637	5
Gamma	0.045	0.944	6
Twin sensitivity	0.043	0.946	6
Mixture	0.046	0.944	5

Note. AAD = Average Absolute Deviation; PVA = Percentage of Variance Accounted for (R^2); FP's = free parameters.

Models of the Experimental Materials

In the 5-4 category structure, individual features of value 1 (assignment of 1 and 0 is arbitrary in a particular experiment) dominate the five A stimuli and features of value 0 dominate the B stimuli: A stimuli average 2.8 1-features; B stimuli average 1.25. It is reasonable that participants would learn that 1-values indicate probable A membership, and 0 values, probable B membership. We refer to this characteristic of stimuli in the 5-4 category structure as "A-proneness."

Suppose we assign to each stimulus a value corresponding to its "A-proneness." For example, we could assign a stimulus 4 points for a 1 in F1, 3 points for a 1 in F2, 4 points for a 1 in F3, and 3 points for a 1 in F4, corresponding to the frequency with which these feature values are associated with category A in the learning set. Then take the sum of these values as the measure of A-proneness for a particular stimulus. By this method, stimulus 4A gets a score of 11, and stimulus 12B gets a score of 7.

We can fit this measure of A-proneness to the Smith and Minda (2000) data (percentage of category A responses for each stimulus) using a linear regression. The result is a 2-parameter model ($y = .034x + .069$) that predicts the percentage of category 'A' responses for all 16 stimuli with an R^2 of .81 and an AAD of .098. This simple linear regression, with only two free parameters, and derived strictly from the feature structure of the stimuli, accounts for the lion's share of the performance variability.

This analysis does not instill confidence that the more elaborate models have much to say about the actual psychological processes of the participants in these experiments. The only thing any of these models (including our strawman model) tell us about participants in these studies is that they are capable of tuning their behavior to the feature structure of the stimuli. These are more models of the experimental materials than they are models of psychological processes.

Indeed, Smith and Minda (2000) observe that the best of the mathematical models all produced such good fits to the data that it was impossible to choose between the very different process models motivating them:

> ". . . the underlying representation and process remains undetermined and unknown. Therefore, one sees that the [29] 5-4 data sets, when described by formal models, are silent on the matter of whether categories are represented in a way that is based on prototypes or in a way that is based on exemplars." (p. 17)

Smith and Minda (2000) conclude from this that the 5-4 category structure is too limited in its properties for general conclusions to be drawn from it about the processes that people use to learn new categories.

We do not agree that the 5-4 category structure is inherently too limited to reveal the underlying cognitive processes. We propose instead that the methods dominating this area of research are too limited. You've got to have the right tool for the job. In this case, the job is to uncover the processes underlying category learning and categorization

performance. We claim that the right tool (methodology) for this job is fine-grained information-processing analyses, using a variety of converging measures. We have found that detailed analysis of the trial-by-trial behaviors of individual participants reveals rich complexity in their categorization processes. In the next section, we describe the completed study, our approach to data analysis, and the lessons we have learned about the complexity and variability of categorization processes in the 5-4 paradigm.

An Information-Processing Analysis

The study we completed involved a partial replication of Medin and Smith (1981), whose category learning study implemented the 5-4 category structure using Brunswik faces. The four binary features of Brunswik faces are Eye Height (EH – High and Low), Eye Spacing (ES – Wide and Narrow), Nose Length (NL – Long and Short), and Mouth Height (MH – High and Low). These features – EH, ES, NL, and MH – correspond to features F1-F4, respectively, in Table 1. Like Medin and Smith, we used three instruction conditions (Standard, Prototype, and Rule-X), a learning phase, and a transfer phase. After that, participants were presented each of the 16 faces and its associated category, one at a time, and they rated the extent to which the face was typical for that category on a 1-to-9 scale. At the end, participants gave retrospective reports describing the processes they used to categorize the stimuli. Thirty-six Carnegie Mellon University undergraduates participated in this study. Half gave concurrent verbal protocols during the entire learning phase. Our analyses focus on these 18 participants.

A Variety of Measures

The data we have focused on in our process analyses include (1) errors, (2) concurrent verbal protocols, (3) typicality ratings, and (4) retrospective reports. Data analysis were not limited to measuring the frequency with which participants choose A or B responses to the 16 stimuli during a transfer trial. Instead, we relied on analysis of the detailed behavior of participants while they were performing both the learning and the transfer task: data that revealed a great deal about the processes they were using.

In performing these analyses, we have been guided by the idiographic data analysis methods typified by Newell and Simon (1972) and by a general theory of perception and memory, EPAM, that has previously been applied to aggregate data on the 5-4 task (Gobet, Richman, Staszewski, & Simon, 1997). EPAM is a computer program that uses a discrimination net architecture to simulate the participants' behavior in responding to each stimulus. In fact, EPAM was used to simulate the aggregate data from Medin and Smith (1981) that comprises three of the 29 Smith and Minda (2000) data sets.

Following are descriptions of our data analysis procedures, accompanied by illustrations of how analysis at this level of detail can reveal participants' categorization processes.

Errors. Participants may use from one to four features to classify a face, and they exercise most of these options at one time or another. Table 3 shows the likely errors that arise (out of ambiguities) when the nine faces used in the learning trials are categorized only on the basis of particular features, or particular pairs, or triplets of features.

The four rows and the first four columns of the table name the features. Each of the cells in the first four columns corresponds to a classification of the faces on the corresponding pair of features. For convenience of reference, we have designated the cells of the table corresponding to particular combinations of feature tests with letters from M through Z.

For example cell V, which is at the intersection of row EH and column NL, shows on the first line that 5A and 2B cannot be distinguished using only these two features, for both faces have identical eye heights and nose lengths (EH=0; NL=1). Similarly (second line), 13A and 12B are identical on EH (1) and NL (0), so a discrimination process that relied only on those two features would not be able to discriminate stimuli 13A and 12B. The other five faces form two classes: A's with high eyes and long noses, and B's with low eyes and short noses. So, for this particular pair of tests, four faces are ambiguous or "hard," and likely to be misclassified during learning. If, instead of EH:NL, the features attended to were nose length and mouth height (NL:MH), then 4A, 7A and 2B would fall in a single class, as would 13A and 14B, and these five would be the hard faces in this case. Thus, when participants use a particular pair of features to classify faces, they will make the greatest number of errors in classifying the faces that are hard for that pair.

The column of Table 11 marked "EXCEPT" indicates which faces would be error-prone if the three features *except* the one labeling the corresponding row were tested (i.e., a discrimination net using the three remaining features); the column marked "SOLE" indicates which faces would be error-prone if *only* the feature on that row were tested. The EXCEPT column shows that all nine faces can be categorized perfectly without the use of ES, but the three other features, EH:NL:MH, must all be used. Notice that each of the three triads of features that includes ES produces a different set of hard faces, as does each net using only a particular single feature.

By assessing which were the hardest faces during the learning phase, we identified the dominant discrimination strategy for each participant. Participants in the Prototype instruction condition showed the most between-subject variability in process, with strategies V, W, Y, P, and R inferred from their errors in the learning phase. Standard participants also showed considerable variability, with evidence of strategies V, W, Y, and Z in their data. The Rule-X participants, who were told explicitly to attend to nose length, used strategies R and V.

Table 3: Error patterns predicted by feature selection in the 5-4 categorization paradigm

	EH (F1)	ES (F2)	NL (F3)	MH (F4)	EXCEPT	SOLE
EH (F1)		5A,2B 4A,13A,12B **U**	5A, 2B 13A,12B **V**	5A,14B 4A,7A,12B **W**	4A,2B **M**	5A, 12B **N**
ES (F2)			4A,5A,2B 13A,12B **X**	15A,14B 7A,10B 4A,2B,12B **Y**	**O**	7A,15A 2B,12B **P**
NL (F3)				4A,7A,2B 13A,14B **Z**	4A,12B **Q**	13A,2B **R**
MH (F4)					13A,12B 5A,2B **S**	4A,7A 14B **T**

Note. Stimuli listed in each cell (e.g., 5A, 12B) are those for which errors are expected if the participant is attending to that conjunction or disjunction of features. Bold code letters (e.g., **U**, **V**, **W**) are used in the text as an economical means of referring to specific categorization strategies, as indicated by increased attention to specific features. F1-F4 = Features 1-4 in Table 1. EH = Eye Height; ES = Eye Spacing; NL = Nose Length; MH = Mouth Height. EXCEPT = attention to all features except the feature in that row. SOLE = attention to only the feature in that row.

Verbal Protocols. We assume that the features used in discriminating and categorizing the faces are verbalizable. The claim is not that participants will verbalize every feature to which they attend, or even that discrimination is always a verbal process; the claim is that the process of encoding features can create a verbalizable representation, and that patterns of verbalization of features are correlated with patterns of attention to the stimuli.

The Brunswik faces are easy to distinguish visually, and to describe verbally, using either the “official” features (eye height, eye spacing, nose length, mouth height) mentioned in the experimental instructions, or other descriptors that may be already familiar to individual participants(e.g., "long face", "small distance between nose and mouth," "wide face," or even "monkey-like"). The official descriptors, rather than idiosyncratic ones, are by far the more frequent in the protocols.

Participants' protocols mainly reported values of features of the face they were currently categorizing, sometimes supplemented with a reason for assigning the face to a particular class, and sometimes with a comparison with a previous face. The following (each preceded by identifier of participant and experimental condition) are examples of verbal responses to stimuli that described features in the language of the instructions:

MS (prototype). "High eyes; short nose; low mouth. Let's go with B, because the last one had high eyes and low mouth."

ML (prototype). "I'll say this is A because of the nose length and the eye height and the separation between the eyes and the mouth."

Rather more austere and more typical are:

MK (standard). "close and high eyes, small nose, middle mouth." *(Chooses A.)*

RB (prototype). "The eyes are low and the nose is big." (*Chooses B*.)

The discrimination processes of participants who use idiosyncratic descriptive terms are harder to identify, but the descriptors they actually used were generally related to the "official" ones in simple ways. For example, "long" faces were faces with high eyes, and sometimes also with low mouths. Faces with "eyes close to the nose" were faces with low, close eyes. The meaning, in terms of features, of these non-standard descriptors can usually be determined by checking the characteristics of the faces to which participants applied them.

Typicality Ratings. Following the transfer phase of the experiment, participants were shown each face along with its correct category. Their task was to rate the extent to which each face was typical for its category. The verbal protocol participants also provided explanations for their ratings. These proved to be informative as additional converging evidence regarding how participants were discriminating the stimuli. Following are several examples from the typicality rating explanations:

JIS (standard). “This one is pretty typical of A because the eyes are way up high and spread out in this one.”

RB (prototype). “That one, I think, is typical because the eyes are high and far apart, and the nose is little.”

MB (rule-x). “Typical. Short nose.”

In addition to lists of features as justification for the ratings, participants occasionally referred to Gestalt characteristics of the faces, using terms like "long" or "wide". For instance, "This one doesn't look like a B face. The eyes are high, and it looks like a kind of long face." The majority of explanations, however, were feature lists that revealed the various ways participants used combinations of feature values to categorize the stimuli.

Retrospective Reports. After the typicality ratings, the experimenter asked each participant, "On what basis were you making your classifications?" Following are two example responses:

RB (prototype). "Most of the type A had high eyes, and it didn't matter where the nose is or the mouth. And most of the B's had eyes in the middle, but there was a type B that had really high eyes. And then there was a type A that had eyes in the middle with a little nose and a long mouth."

MK (standard). "Basically, small nose was A, big nose was B. Basically, except small nose if the mouth was low, I looked at the eyes, and if the eyes were low, then it was B. If it was a big nose with little mouth and high up, then I checked the eyes, and if the eyes were high, then the face was A."

Note that both RB's and MK's retrospective reports are consistent with their concurrent verbalizations from the learning trials. It is converging evidence of this sort that increases our confidence in conclusions regarding participants' categorization processes.

Comparison of VP and NVP Errors

An assumption that is required in drawing generalizations from the verbal protocol participants is that the requirement to give protocols does not itself have a direct impact on categorization processes in this task. It would be reassuring if the performance of the verbal protocol (VP) participants and the non-verbal protocol (NVP) participants were in fact similar.

Following the logic in Table 3, to the extent participants in the VP and NVP conditions found similar faces difficult during the learning phase, there is evidence for similar categorization strategies across those conditions.

In both conditions, the four most difficult stimuli are 2, 5, 12, and 13. These are difficult stimuli because they are exceptions on features that are highly predictive of category membership. Table 1 shows that stimuli 2 and 13 are exceptions on Nose Length, while 5 and 12 are exceptions on Eye Height. Additionally, pairs of those stimuli are confusable if one ignores Mouth Height. That is, 2 and 5 are identical except for mouth height, as are 12 and 13. The fact that these four stimuli are always the most difficult, suggests that most participants, regardless of verbal protocol condition, found it difficult to learn the exceptions.

Looking at error rates across all of the stimuli, we find that VP and NVP error rates correlate $r = .82$, indicating a high degree of similarity in the error patterns between the two conditions. In terms of overall error rate, VP participants tended to make more errors (Mean = 50.2) than NVP participants (Mean = 39.7), although an ANOVA reveals that this difference is not significant: $F(1, 36) = 1.219$, $p = .277$).

Summary of Findings

Due to space limitations, and because of the massiveness and complexity of the body of data we are examining, we must briefly summarize our most important findings. Additionally, we feel that a general description of the results will be more useful, in terms of distinguishing the information processing approach from the more typical, aggregate-level, nomothetic approach, than the specific frequencies and percentages in our findings would be. Therefore, we will finish the paper with an account of our main findings.

One lesson that emerges from these rich data is that the task structure itself was a major determinant of the outcomes we measured. The influence of task structure on performance is apparent in Figure 2, and we found similar effects in our data. Because nearly all of our participants achieved the learning criterion, we infer that they discovered the implicit task structure.

A second finding is that most of the participants, regardless of their instructional condition, interpreted the task as one of forming rules that could be used to assign faces to a category. This generally took the form of learning what *features* were associated with the A or B categories, then using this knowledge to classify *faces* by means of their features, rather than defining prototypic faces or cycling through comparisons with previous exemplars. Feature-based rule following is apparent in the verbal protocols presented earlier.

Regarding rule-based behaviors, some (but not all) participants discovered that it was also useful to recognize certain individual faces and associate their categories directly with them. These were almost always the "hard" faces that did not fit the simple rules they used to discriminate the others. This was particularly apparent in statements like, "Ah, this is the one that tricks me." This phenomenon is consistent with the model of Nosofsky, Palmeri, and McKinley (1994), but the inconsistent appearance of this phenomenon in our data also suggests that model's limitations.

Another finding is that the numerous differences in the behaviors of different participants could be traced in large measure to different strategies of attention, and different strategies for retaining and combining information about features and combinations of features upon which rules of choice could be built. Strategies were effective to the extent that they made only modest demands on memory, including demands on short-term memory and demands for transferring information to long-term memory and retaining it for use in building up the structure of decision rules.

Perhaps the most consistent phenomenon we observed in the data was the high degree of within-participant variability in process. Examination of participants' stimulus-by-stimulus category responses and corresponding

verbalizations reveals that the dominant rule-based processes are sprinkled with instances of comparing the current stimulus to the immediately preceding stimulus, and also increasing evidence of recognition-based processes (especially for the hard faces) with experience. Even within the rule-based processes, there was a good deal of variability, as participants had varying degrees of success with the feature-based rules they generated and tested.

Conclusion

Sciences are concerned with discovering and testing laws that describe the invariant features of their domains. Invariance is a complex concept. Even the gravitational constant is not an invariant once one strays from the Earth or ascends a mountain. So science has laws, like the law of gravitational attraction, but it has parameters and variables that specify the workings of each law as a function of various circumstances.

Matters become especially complex when we consider laws of biology, with its immense variety of living forms, and still more complex when we consider the laws of psychology, which seeks the regularities in the behavior of an organism that has enormous capabilities for adaptation and learning. Not only will the behavior vary with the innumerable features of the environment in which the person performs the task, but even leaving genetic differences aside, it will vary as a function of each individual's previous history of experience and instruction.

The experimental data we have analyzed here illustrate a number of such complexities. No unitary set of laws, taken by itself, governs the precise way in which a set of people go about solving a simple categorization task, not even if all of them are drawn from the same university population. Covering variation by averaging conceals it but does not banish it or explain it. Siegler (1987) made the same point about the "Perils of Averaging," but in the context of analyzing children's arithmetic. With this paper we illustrate the value of a fine-grained, multivariate approach for advancing our understanding of categorization and category learning in terms of their underlying cognitive processes.

It was almost a half century ago that Bruner, Goodnow, and Austin (1956) published their seminal work on categorization, *A Study of Thinking*. In the introduction to that book, they wrote:

> ". . . we have come gradually to the conclusion that what is most needed in the analysis of categorizing phenomena . . . is an adequate analytic description of the actual behavior that goes on when a person learns how to use defining cues as a basis for grouping the events of his environment." (p. 23)

To understand participants' actual behaviors during categorization and category learning processes, not only did we need to analyze the behavior of individual participants, but the data obtained from each had to be of a grain size fine enough to capture some detail of ongoing learning processes. We examined errors and concurrent verbalizations stimulus-by-stimulus, then looked for converging evidence in participants' typicality ratings and their retrospective reports. The requirement that conclusions about participants' processes should be based on the convergence of multiple measures provided strong tests of the validity of our findings. In the end, our approach yielded a rich, descriptive understanding of the underlying representations and processes employed by participants in our study.

The challenge to explain the phenomena observed remains. Because of the variety of measures used in our analyses, and the variation among them in granularity, we advocate the development of simulation models. We submit that detailed, multivariate information-processing analyses and simulation modeling are tools that are well-suited for the job of advancing understanding of category learning and categorization. The data help us come to a better understanding of actual cognitive processes in category learning, and simulation models allow for the possibility of accounting for the enormous variability in process within and between participants.

Acknowledgements

The authors would like to thank Damian Bierman for assisting in the early stages of this research project. The opinions expressed in this paper are those of the authors and do not reflect the official policy or position of the United States Air Force, the U.S. Department of Defense, or the U.S. Government.

References

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.

Gobet, F., Richman, H. B., Staszewski, J., and Simon, H. A. (1997).Goals, representations, and strategies in a concept formation task: The EPAM model. In *The Psychology of Learning and Motivation, Vol. 37*, D. L. Medin (Ed.), pp. 265-290. San Diego, CA: Academic Press.

Medin, D. L., & Smith, E. E. (1981). Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7*(4), 241-253.

Newell, A. & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nosofsky, R. M., Palmeri, T. J., &McKinley, S. C. (1994). Rule-plus-exception model of category learning. *Psychological Review*, *101*, 53-79.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General, 116*(3), 250-264.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(1), 3-27.

Is Experts' Knowledge Modular?

Fernand Gobet (frg@psyc.nott.ac.uk)
ESRC CREDIT
School of Psychology
University of Nottingham
Nottingham NG7 2RD, UK

Abstract

This paper explores, both with empirical data and with computer simulations, the extent to which modularity characterises experts' knowledge. We discuss a replication of Chase and Simon's (1973) classic method of identifying 'chunks', i.e., perceptual patterns stored in memory and used as units. This method uses data about the placement of pairs of items in a memory task and consists of comparing latencies between these items and the number and type of relations they share. We then compare the human data with simulations carried out with CHREST, a computer model of perception and memory. We show that the model, based upon the acquisition of a large number of chunks, accounts for the human data well. This is taken as evidence that human knowledge is organised in a modular fashion.

Introduction

An important goal of cognitive science is to understand the characteristics of knowledge, in particular the way it is acquired and used. To achieve this goal, research has employed a number of methods, including artificial laboratory experiments, such as nonsense syllable learning, and collection of naturalistic data, such as experts functioning in their natural environments. It is generally accepted that knowledge consists of different types (declarative, procedural, episodic) and that its acquisition follows a power law of learning. In addition, it has often been proposed that knowledge is modular, consisting, for example, of productions (e.g., Newell, 1990) or of perceptual chunks (e.g., Chase & Simon, 1973).

The goal of this paper is to explore, both with empirical data and with computer simulations, the extent to which modularity characterises human knowledge, and in particular experts' knowledge. We first describe the concept of modularity, and then show how it has been used in expertise research. This leads us to describe the CHREST architecture, which acquires knowledge by growing a discrimination net encoding chunks. Next, we present data aimed at characterising the properties of experts' chunks, and compare them with those acquired by CHREST. The comparison results in an excellent fit between the model and the human data. In the conclusion, implications for the modularity of human knowledge in general are drawn.

Modularity of Knowledge

Several formalisms, both modular and non-modular, have been developed in cognitive science to explain how humans represent and implement knowledge. Examples of modular representations are production systems, semantic networks, and discrimination nets. Examples of non-modular representations are distributed neural networks, holograms, and various mathematical representations based on matrix algebra. This classification should be considered with caution, however. On the one hand, production rules, for example, are typically organised in problem spaces (e.g., Newell, 1990), and their interdependence can be considerable, which counts against strict modularity. On the other hand, it could be argued that, in non-modular representations, modules emerge as the system develops or learns (e.g., Rumelhart & McClelland, 1986).

Modular knowledge organisation has attracted much interest in computer science and artificial intelligence, given the importance of how knowledge is indexed, structured, organised, and retrieved (e.g., Lane et al., 2000). In artificial intelligence, modularity has been defined as "the ability to add, modify, or delete individual data structures more or less independently of the remainder of the database, that is, with clearly circumscribed effects on what the system 'knows' " (Barr & Feigenbaum, 1989, p. 149). While a strong argument can be made that it is easier to understand modular and decomposable systems than systems that do not share these properties (e.g., Simon, 1969), and that the value of these properties has been demonstrated in fields such as software engineering, it is an empirical question whether human knowledge is modular or not. A rich source of data about this question has been gained from research into expert behaviour, to which we now turn our attention.

Chess Experts' Knowledge

In his seminal study, De Groot (1946/1965) subjected chess players to a number of problem-solving and memory experiments. The surprising result was that, in a choice-of-a-move task, there was no large skill difference in variables such as depth of search, number of moves considered, or search heuristics employed. However, a clear difference was found in a memory task where a chess position was presented for a few seconds. Masters could recall the entire position almost perfectly,

while weaker players could recall only a handful of pieces. De Groot concluded that expertise does not reside in any superior abilities but in knowledge.

Continuing de Groot's research, Chase and Simon (1973) carried out a study destined to have a huge impact in cognitive science. They used two tasks. In the *recall task,* based on de Groot's (1965) method, a chess position was presented for five seconds, and players had to reconstruct as many pieces as possible. In the *copy task,* the stimulus board remained in view, and the goal was to reconstruct it onto a second, empty board. As the stimulus and the reconstruction boards could not be fixated simultaneously, Chase and Simon used the glances between the boards to detect memory chunks. Comparing the latencies between successive pieces in the copy and recall tasks, they inferred that pieces replaced with less than 2 seconds' interval belonged to the same chunk, and that pieces placed with an interval of more than 2 seconds belonged to different chunks. Finally, they showed that the chunk definition based upon the latencies between two successive pieces was consistent with a definition based upon the pattern of semantic relations (attack, defence, proximity, colour, and type of piece) shared by these two pieces. This converging evidence was used to infer the chunks used to mediate superior performance, and to explore how they allowed masters to find good moves despite their highly selective search. A number of other experimental tasks (reviewed in Gobet & Simon, 1998) have brought converging evidence for the psychological reality of chunks, as defined either by latency in placement or by number of relations between pieces.

Simon and Gilmartin (1973) developed a computer program (MAPP; Memory-Aided Pattern Perceiver) implementing some of Chase and Simon's ideas. MAPP is based upon EPAM (Elementary Perceiver and Memorizer; Feigenbaum & Simon, 1984), a theory developed to account for empirical phenomena where chunking (i.e., acquisition of perceptual units of increasing size) is seen as essential. The basic idea in MAPP was that long-term memory (LTM) is accessed through a discrimination net, and that, once elicited, LTM chunks are stored in short-term memory (STM) through a pointer. MAPP's relatively low recall performance—slightly better than a good amateur, but inferior to an expert—was attributed to the small number of nodes, about two thousand, stored in its LTM. MAPP simulated several results successfully: increase in performance as a function of the number of chunks in LTM; kind of pieces replaced; and contents of chunks. However, in addition to its failure in simulating expert behaviour, the program had several limitations (De Groot & Gobet, 1996). In particular, the chunks were chosen by the programmers and not autonomously learnt, and the program made incorrect predictions for a number of experiments that were later carried out. These limitations were removed in the CHREST program discussed below.

CHREST

CHREST (Chunk Hierarchy and REtrieval STructures; De Groot & Gobet, 1996; Gobet & Simon, 2000) is a cognitive architecture similar to MAPP. CHREST originally addressed high-level perception, learning and memory, but various problem-solving mechanisms have been implemented recently. It is composed of processes for acquiring low-level perceptual information, an STM, attentional mechanisms, a discrimination net for indexing items in LTM, and mechanisms for making associations in LTM such as production rules or schemas. STM mediates the flow of information processing between the model's components. The central processing of CHREST revolves around the acquisition of a discrimination net based on high-level perceptual features picked up by attentional mechanisms and on the creation of links connecting nodes of this net together.

After the simulated eye has fixated on an object, features are extracted and processed in the discrimination net, and then, based upon the output of the discrimination, a further eye fixation is made, and so on. STM operates as a queue; that is, the first elements to enter are also the first to leave. STM has a limited capacity, which consists of four chunks (Cowan, 2001; Gobet & Simon, 2000). Processing is constrained by a number of restrictions, including time parameters such as the time to fixate a chunk in LTM (8 s) and capacity parameters such as the four-chunk limit of STM.

The discrimination net consists of *nodes*, which contain *images* (i.e., the internal representation of the external objects; images correspond to Chase and Simon's *chunks*); the nodes are interconnected by *links*, which contain *tests* allowing items to be sorted through the net. Learning happens as follows: once an item has been sorted through the net, it is compared to the image in the node reached. If the item and image agree but there is more information in the item than the image, then *familiarisation* occurs, in which further information from the item is added to the image. If the item and image disagree in some feature, then *discrimination* occurs, in which a new node and a new link are added to the net. Based on empirical data, it has been estimated that discrimination requires about 8 s and familiarisation about 2 s.

In addition to these learning mechanisms, CHREST has mechanisms for augmenting semantic memory by the creation of schemas (known as *templates*) and of *lateral* links connecting nodes together (Gobet, 1996); for example, these links can be created when nodes are sufficiently similar ('similarity links'), or when one node can act as the condition of another node ('condition links'). The creation of these links is consistent with the emphasis on processing limits present in both EPAM and CHREST, in that all nodes used for creating new links must be in STM.

Table 1: Copy, recall and *a priori* chess relations probabilities, for combinations of the five chess relations: Attack (A), Defence (D), Spatial Proximity (P), Same Colour (C), and Same Piece (S).

	COPY				**RECALL**				***A priori* Probabilities**	
Relations	GAME		RANDOM		GAME		RANDOM		GAME	RANDOM
	WITHIN	BETWEEN	WITHIN	BETWEEN	≤ 2 sec	> 2 sec	≤ 2 sec	> 2 sec		
-	.037**	.172**	.086**	.129**	.052**	.190**	.051**	.284	.335	.297
A	.005**	.006	.031	.054**	.004**	.024	.000*	.054	.016	.024
P	.000	.006	.037**	.059**	.001	.006	.033**	.041*	.004	.010
C	.148**	.278	.152**	.203**	.132**	.247	.136**	.189	.255	.297
S	.016**	.056**	.040**	.049**	.040**	.102*	.059**	.054	.154	.144
AP	.000*	.000	.056**	.069**	.001	.003	.015	.027	.005	.028
AS	.000	.000	.003	.005*	.004**	.003	.000	.000	.001	.001
DC	.104**	.133**	.072**	.077**	.059**	.084**	.044	.068*	.035	.024
PC	.084**	.067**	.059**	.046**	.049**	.060**	.066**	.081**	.019	.009
PS	.002	.006	.044**	.064**	.006	.012	.018	.027	.006	.010
CS	.115	.094	.135*	.105	.111	.057*	.059*	.041	.096	.108
APS	.000	.000	.013*	.013	.001	.000	.018	.014	.001	.007
DPC	.109**	.078	.123**	.064**	.093**	.084*	.118**	.081*	.048	.028
DCS	.048**	.017**	.000	.000	.033**	.012**	.015**	.000	.002	.001
PCS	.196**	.039**	.127**	.039**	.202**	.060**	.232**	.041**	.011	.007
DPCS	.137**	.050**	.023**	.023**	.213**	.054**	.136**	.000	.013	.007
#observations	1283	180	1114	389	1563	332	272	74		

Note: * means p<.01, ** means p<.001 (both two-tailed). The statistical significance levels are based on the *z*-values that were computed using the following formula (assuming the normal approximation to the binomial distribution):

$$z = \frac{p_o - p_e}{\text{s. e.}}, \qquad \text{where s. e.} = \sqrt{\frac{p_e\,(1 - p_e)}{\text{sample size}}},$$

and where p_o is the observed probability and p_e the *a priori* (expected) probability.

CHREST can reproduce a number of features of the behaviour of skilled and unskilled chess players in memory experiments, such as their eye movements, the size and number of chunks, the number and type of errors, and the differential recall of game and random positions (De Groot & Gobet, 1996; Gobet & Simon, 2000). As a psychological theory, CHREST has several strengths. It is parsimonious, with few free parameters. It provides absolute quantitative predictions, for example about the number of errors committed or the time taken by a subject to carry out a task. Together with EPAM, it simulates in detail a number of empirical phenomena from various domains, such as verbal learning, context effects in letter perception, concept formation, expert behaviour, acquisition of first language by children, and use of multiple representations in physics (see Gobet et al., in press, for a review).

A Replication of Chase and Simon (1973)

As noted above, Chase and Simon (1973) operationalised the concept of chunk using both the latencies between successive piece placements and the semantic relations between them. Their experiment has recently been replicated and extended by Gobet and Simon (1998). The main difference between the two studies is that Gobet and Simon used a computer display to present the tasks instead of physical chessboards. In spite of this difference, there is an important overlap between the results of the two studies.

Gobet and Simon analysed 26 players (Chase and Simon had only 3) ranging from good amateurs to professional grandmasters, who were divided into three skill levels (Masters, Experts and Class A players). The results were in line with previous experiments, showing a massive skill effect with game position, and a small but reliable skill effect even with meaningless positions. Here, we focus upon the operationalisation of chunks, relying both upon Gobet and Simon's published data and upon additional analyses.

Latencies Predict Chunk Boundaries

Gobet and Simon essentially followed Chase and Simon's approach. They first estimated a time threshold (2 s) as a means to decide whether two pieces placed in succession belonged to the same chunk, and then

validated this threshold by showing that it led, on average, to similar chunks as those obtained by using semantic relations. If they are modular, chunks should be characterised by a high density of relations between the elements that constitute it, and by a low density of relations with elements from other chunks (Chase & Simon, 1973; Cowan, 2001). That is, there should be many more relations between successive pieces within the same chunk than between successive pieces on opposite sides of a chunk boundary. Thus, the relations between successively replaced pieces should be different depending on whether they are separated by short or long latencies. In addition, assuming that the same cognitive mechanisms mediate the latencies in the copy and recall experiments, the two experiments should show the same pattern of interaction between latencies and number of relations. In other words, the relations for the within-glance placements in the copy task should correlate with those for rapid placements (≤ 2 s) in the recall task and the relations for between-glance placements in the former should correlate with those for slow placements (> 2 s), in the latter.

These predictions are met in both the copy and the recall tasks, whose results correlate highly. Within chunks, small latencies correlate with a large number of relations, while large latencies occur when there are few relations between successive pieces. No such relationship is observed for successive pieces belonging to different chunks. The shortest latencies are found with four relations (Defence, Proximity, Colour, and Kind), which mainly occur with pawn formations.

Relations Predict Chunk Boundaries

The next step consists in showing that the pattern of relation probabilities for within-chunk, but not for between-chunk placements, differs from what could be expected by chance. Table 1 gives the probabilities of the presence of different combinations of relations in the various experimental conditions, with the three skill levels pooled. The last two columns give the *a priori* probabilities (for game and random positions, respectively) that were calculated by recording, for each position, all relations that exist between all possible pairs of pieces; the *a priori* probability for a relation is obtained by dividing the total number of occurrences of a relation by the total number of possible pairs. These *a priori* probabilities were based on 100 positions and 26,801 pairs. Finally, the *z*-values indicate whether the observed probabilities reliably differ from the *a priori* probabilities.

In the copy task, with game positions but not with random positions,[1] the between-glance probabilities are much closer to chance than the within-glance probabilities. This pattern holds also in the recall of both random and game positions when slow placements (> 2 s) are compared with fast placements (≤ 2 s). The probabilities for pieces with three and four relations are high in the within-glance and fast (≤ 2 s) conditions compared with the between-glance and slow (> 2 s) conditions; the opposite is true for pieces with one relation or none. Note also that the probabilities for combinations of relations that include an attack (A) are conspicuously low, compared with chance, for game positions but not for random positions.

One way to make sense of Table 1 is to analyse the correspondence between the number of chess relations and the deviations from *a priori* probabilities, computed by subtracting the *a priori* probabilities from the observed frequencies of a given condition. Based on the notion of modularity, it should be expected that the within-chunk deviations from *a priori* probabilities would be highly correlated with the number of relations, while this would not be the case for the between-chunk deviations. This is exactly what was found. The correlations with number of relations are high for the within-chunk conditions (copy game within-glance: 0.81; copy random within-glance: 0.68; recall game short latencies: 0.86; recall random short latencies: 0.79; all the correlations are statistically significant at $p = .005$). The correlations are smaller with the between-chunk conditions (copy game between-glance: 0.61; copy random between-glance: 0.56; recall game long latencies: 0.58; recall random long latencies: -0.15; none of the correlations are significant at the .01 level). These results are illustrated graphically in Figure 1, which shows the results for game and random positions as a function of whether the placements were within-chunk or between-chunk. From the Figure, it is clear that, for within-chunk conditions, the placements having few relations are below chance, while the placements having several relations are above chance. There is no such clear relation for the between-chunks placements.

Computer Simulations

We now show that CHREST captures the composition of chunks and the pattern of relations of within- and between-chunk placements. Simulations of similar phenomena, carried out by Simon and Gilmartin (1973) using MAPP, were limited to a single subject and matched the data only approximately.

Methods

In the *learning phase*, the program scanned a large database of master-game positions, fixating squares with simulated eye movements, and learning chunks using discrimination and familiarisation. Three nets were created, estimated to correspond roughly to the recall percentages of Class A players, experts, and masters with a five-second presentation time. These nets had respectively 1,000 nodes, 10,000 nodes, and 100,000 nodes.

For the simulations of the *performance phase*, the program was tested with 100 game positions and 100

[1] That this pattern does not hold with the copy of random positions may be due to the strategy used by subjects to replace these positions. Several subjects copied the positions line by line or column by column.

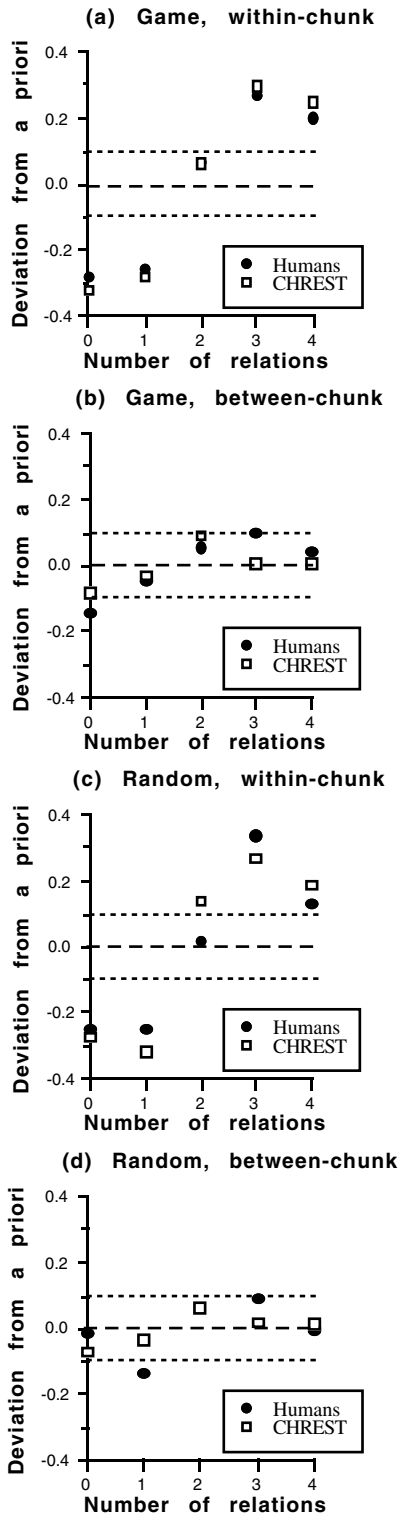


Figure 1: Relation between chess relation probabilities and the number of relations shared by two pieces successively placed. The long-dash line indicates zero deviation, and the short-dash lines indicate deviations of 0.1 above or below zero.

random positions. Learning was turned off. During the five-second presentation of a position, CHREST moved its simulated eye around the board. Each eye fixation defined a visual field (all squares within two squares from the square fixated); the pieces within the visual field are treated as a single pattern and sorted through the discrimination net. Other patterns are defined by the pieces focused upon in two successive eye fixations. If a chunk is found in the discrimination net, a pointer to it is placed in STM.

During the reconstruction of a position, CHREST used the information stored in STM. When a piece belonged to several chunks, it was replaced only once. In case of conflicts (e.g., a square is proposed to contain several pieces), CHREST resolved them sequentially, based on the frequency with which each placement is proposed. Like humans, it sometimes made several different proposals about the location of a piece or about the contents of a square. Finally, some weak heuristics were used, such as the fact that only one white king can be replaced in a position. (See Gobet & Simon, 2000, for more detail.)

A chunk refers to the image of a node in the discrimination net. It is therefore straightforward to decide whether two pieces do or do not belong to the same chunk. The relations between pieces were extracted using the same program as that used with the human data.

Results

Table 2 gives the probabilities of observing a pattern of relations, as a function of the type of position and the kind of placement. Although the fit with the corresponding human data shown in Table 1 is reasonable

Table 2. Recall and *a priori* chess relations probabilities, for combinations of the five chess relations: Attack (A), Defence (D), Spatial Proximity (P), Same Colour (C), and Same Piece (S).

	Game positions			Random positions		
Relations	Within	Between	*A priori*	Within	Between	*A priori*
-	.009	.254	.335	.018	.231	.297
A	.005	.034	.016	.021	.061	.024
P	.013	.011	.004	.050	.026	.010
C	.104	.208	.255	.040	.216	.297
S	.021	.148	.154	.050	.136	.144
AP	.004	.013	.005	.030	.027	.028
AS	.000	.001	.001	.001	.005	.001
DC	.042	.059	.035	.038	.042	.024
PC	.097	.050	.019	.092	.039	.009
PS	.020	.019	.006	.061	.018	.010
CS	.064	.113	.096	.094	.111	.108
APS	.004	.005	.001	.008	.017	.007
DPC	.162	.031	.048	.148	.033	.028
DCS	.007	.000	.002	.009	.001	.001
PCS	.186	.032	.011	.147	.015	.007
DPCS	.259	.021	.013	.193	.023	.007

(the r^2 are: game within-chunk: .83; game between-chunk: .82; random within-chunk: .58; random between-chunk: .75), not too much weight should be given to them, because they are sensitive to a few large values, and because they may in part reflect the statistics of the chess environment (i.e., the *a priori* probabilities). As with the human data, we subtracted the *a priori* probabilities from the recall probabilities, and took the sum for each number of relations. Figure 1 shows the results for both the humans and CHREST. The model fits the human data quite well. In particular, the between-chunk placements show little deviation from the *a priori* probabilities, in contrast to the within-chunk placements, which are clearly below chance with zero and one relation, and above chance with three and four relations. All conditions pooled, CHREST accounts for 90% of the variance of the human data.

Conclusion

EPAM and CHREST's learning mechanisms, based upon the construction of a discrimination net of chunks, offer a crisp and computational definition of the concept of knowledge module. Using this definition, Chase and Simon (1973) have found, and Gobet and Simon (1998) have confirmed, that relations and latencies between pieces offer converging evidence for validating the psychological reality of chunks. This paper has shown that, with the same mechanisms used to account for a variety of chess data, CHREST acquires chunks that have the same relational properties as humans'.

The acquisition mechanisms consisting in learning pieces within the visual field and between two eye fixations largely explain the high number of relations within chunks. It is important to note that this phenomenon is not trivial to simulate, however. For example, learning mechanisms such as Saariluoma and Laine's (2001) frequency-based heuristic, where chunk construction is not constrained by spatial contiguity, would fail to account for the data, because they do not capture the relation of proximity which is essential in the chunks acquired by humans (cf. Table 1).

These results, as well as others, indicate that the modular structure of the type of discrimination net used by EPAM and CHREST captures essential aspects of human cognition. Chunks, whose elements share a number of relations, are built up gradually and recursively, with later chunks being built from smaller 'sub-chunks'. Some of these chunks evolve into schema-like structures, and some get later connected by lateral links, thereby constructing both a net of productions and a semantic network. It is not only the presence of a node storing a piece of knowledge which matters, but also the richness with which this node is perceptually indexed and the density with which this node is connected to other nodes. These two aspects give some computational meaning to "conceptual understanding": a richly-connected network of links connecting productions and schemas, that is accessible through perceptual chunks. In addition to expert behaviour, CHREST, which incorporates mechanisms for all these kinds of learning, including the acquisition of modular structures, accounts for empirical phenomena in a variety of domains.

Acknowledgements

I am grateful to Herb Simon for his involvement in many aspects of this research and to the members of the CHREST group for comments on this paper.

Reference List

Barr, A., & Feigenbaum, E. A. (1989). *The handbook of artificial intelligence*. (Vol. 1). New York: Addison-Wesley.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*, 55-81.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*.

de Groot, A. D. (1965). *Thought and choice in chess*. The Hague: Mouton. (First edition in Dutch, 1946).

de Groot, A. D., & Gobet, F. (1996). *Perception and memory in chess.* Assen: Van Gorcum.

Feigenbaum, E. A., & Simon, H. A. (1984). EPAM-like models of recognition and learning. *Cognitive Science, 8*, 305-336.

Gobet, F. (1996). Discrimination nets, production systems and semantic networks: Elements of a unified framework. *Proceedings of the 2nd International Conference on the Learning Sciences* (pp. 398-403). Evanston Il: Northwestern University.

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P.C-H., Jones, G., Oliver, I., & Pine, J. M. (in press). Chunking mechanisms in human learning. *Trends in Cognitive Science*.

Gobet, F., & Simon, H. A. (1998). Expert chess memory: Revisiting the chunking hypothesis. *Memory, 6*, 225-255.

Gobet, F., & Simon, H. A. (2000). Five seconds or sixty? Presentation time in expert memory. *Cognitive Science,* 24, 651-682.

Lane, P. C. R., Gobet, F. , & Cheng, P. C-H. (2000). Learning-based constraints on schemata. *Proceedings of the 22nd Meeting of the Cognitive Science Society*. (pp. 776-782). Mahwah, NJ: Erlbaum.

Newell, A. (1990). *Unified theories of cognition.* Cambridge, MA: Harvard University Press.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing.* Cambridge, MA: MIT Press.

Saariluoma, P. & Laine, T. (2001). Novice construction of chess memory. *Scandinavian Journal of Psychology, 42,* 137-147.

Simon, H. A. (1969). *The sciences of the artificial*. Cambridge: MIT Press.

Simon, H. A., & Gilmartin, K. (1973). A simulation of memory for chess positions. *Cognitive Psychology, 5*, 29-46.

Strategies in Analogous Planning Cases

Andrew S. Gordon (asgordon@us.ibm.com)
IBM TJ Watson Research Center, 30 Saw Mill River Road
Hawthorne, NY 10532 USA

Abstract

Strategies are abstract patterns of planning behavior that are easily recognized, compared, and used by people in everyday planning situations. In an effort to better understand them as a type of mental knowledge structure, three hundred and seventy-two strategies were identified from ten different planning domains, and each was represented in a preformal manner intended to describe the common characteristics of their instances. In doing this large-scale representation work, two observations were made with significance to current theories of analogical reasoning. First, strategies are portions of the relational structure shared by analogous planning cases. Second, representations of strategies include propositions about the reasoning processes of the agent employing them. We propose that a theoretical understanding of analogical reasoning allows us to use strategies as an investigative lens to view the mental models that people have of others and of themselves.

Introduction: Strategy Representations

The term *strategy* periodically appears in cognitive science literature to refer to the abstract patterns that can be recognized in planning behavior. People appear to have a near-effortless ability to use and reason about strategies despite their complexities - as when considering the case of a corporation that underprices part of its product line to bankrupt their competitors, or a case where a parent bird pretends to be wounded to lure a predator away from a place where their nest would be discovered. From the cognitivist's perspective, strategies can be viewed as knowledge schemas, with the assumption that these schemas are mental representations that can be manipulated, compared, and used in intelligent planning and problem-solving behavior. In order to understand this view more deeply, we undertook a project to systematically analyze and represent strategies on a large scale.

We began this project by identifying three hundred and seventy-two strategies from ten diverse domains of planning and problem-solving. Three competitive planning domains were examined: business, warfare, and dictatorship. Three cooperative planning domains were examined: scientific research, education, and personal relationships. Two individual performance domains were included: artistic performance and object counting. Finally, two anthropomorphic domains, where non-people are viewed as planners, were studied: immunology and animal behavior. Strategies for each of these domains were collected using a variety of methods, which included formal interviews with subject-matter experts, introspection, and the analysis of texts such as Niccolo Machiavelli's *The Prince* and Sun Tzu's *The Art of War*, which are nearly encyclopedic of strategies in the domains of dictatorship and warfare, respectively.

For each of these strategies, we authored a definition in the form of a representation such that all situations that match the definition would be positive examples of the strategy, and all cases that do not match the definition would not be examples of the strategy. Recognizing that the same strategy could be applicable in a wide variety of situations - even those that cross domain boundaries - our efforts focused on strategy representations that were of the highest possible level of abstraction while still meeting these definition requirements.

While some descriptive and functional planning languages are beginning to emerge in the artificial intelligence planning community (Tate, 1998; Gil & Blythe, 2000; McDermott, 2000), we chose not to attempt to use them for this representation work, following the belief that these current efforts are not yet expressive enough to describe the subtle planning features found in strategies. Instead, we adopted a style that can be best viewed as *preformal*, somewhat similar to the strategy representations found in smaller-scale efforts (Collins, 1986; Jones, 1992), and where the content of these representations was loosely drawn from a wide range of content theories of planning, notably from Owens (1990). The motivation for using this preformal style was to enable the scaling-up of representation work by relaxing the syntactic formality of logic while preserving the unambiguity of representational terms.

Figure 1 gives three examples of the 372 preformal representations that were authored. Words and phrases in the representations meant to refer to planning concepts and abstractions were capitalized and italicized as they were authored, which allowed us to algorithmically extract them so that they could be analyzed outside of the context of specific

Education strategy: Pop quiz. Ensure that students do their homework with the threat of a unannounced quiz

Representation: The planner has the goal that an agent achieve a set of *Knowledge goals* and has a *Plan* with a *Plan duration* that includes the *Agency* of the agent to execute *Subplans* that are not *Observed executions* by the planner. The planner *Envisions a threat* to the plan in that the agent do not execute the *Subplan* with cause that it is not a *Subplan* of a goal of the agent, and the planner *Envisions the possibility* that the *Successful execution of the plan* would not achieve the goal. The planner *Modifies the plan* by *Scheduling* a set of *Subplans* at *Start times* that are *Randomly selected* from *Moments* in the *Duration*. In each subplan the planner *Requests the execution* of a plan by the agent such that the *Successful execution* of the plan *Requires* that the agent *Executed the subplan* in the goal that were *Subplans previously scheduled*, and where a *Failed execution* will cause a *Violation of a goal* of the agent. The planner adds a *Planning constraint* against plans that cause the agent to have *Knowledge* of the *Scheduled start time* of these subplans.

Animal behavior strategy: Mark your territory. Leave scent marking to avoid unneccessary defensive conflicts

Representation: The planner has a *Competitive relationship* with a set of agents, and has a *Competitive plan* that includes the *Monitoring of execution* of plans of the other agents for *Locations of execution* that are *Locations* in a *Region*, and *In this case* the execution of an *Attack* on the agent. A *Threat* exists that the execution of the *Attack* will be a *Failed execution* with a cause of a *Successfully executed counterplan*. The planner envisions that *If it were the case* that agents in the set had *Knowledge* of the *Competitive plan* of the planner, then a subset of the agents would add the *Planning preference* against plans that had a *Location of execution* in the *Region*. The planner executes a *Repetitive plan* to *Encode information* that is the *Competitive plan* of the planner and *Transfer locations* of the *Encoding of information* to a *Location* that is in the *Region*, where the *Location* is *Selected from the set* with a *Selection criteria* of *Random choice*.

Counting strategy: Transfer between spaces. Count objects as they are moved into an empty location

Representation: The planner has the *Knowledge goal* of the *Quantity* of *Physical objects* in a set. There exists a set of two *Disjoint regions*, where every object has a *Location* that is *Contained within a region* that is a member of the set. The planner has a *Subplan* to *Transfer the location* of a specific object that is *Contained within a region* of the set to a different *Location* that is *Contained within the other region*. The planner executes a plan to achieve the goal that all of the objects in the set have a *Location* that is *Contained within the region* that is the *Start location* in the *Transfer of location* subplan. The planner then *Repetitively executes* a subplan where the planner executes the *Transfer of location* subplan and *Imagines a number*. In the *First repetition*, the number is 1, and in *Subsequent repetitions* the number is the addition of 1 to the *Imagined number* in the *Previous iteration*. The *Termination of repetition condition* is that the planner has an *Execution failure* of the subplan with a cause of *Unfound object in start location*. The planner then *Achieves* the *Knowledge goal* that is the *Imagined number* in the *Last repetition*.

Figure 1. Three preformal representations of strategies from different planning domains

representations. In all, 8,844 italicized words and phrases were extracted from the representations, which was reduced to a set of 974 terms by removing duplicate instances, selecting a representative lexical form for sets of instances that differed only in their inflection, and combining the forms that we determined to be synonymous, i.e. referring to the same planning concept.

The driving motivation behind this large-scale representation work was twofold. First, we aimed to identify the broad representational requirements of strategic planning and outline the mental models that people have of intentional agents. The findings in reference to this first motivation are reported in a separate publication (Gordon, 2001). Our second motivation, which is the subject of this current report, was to further our understanding of the peculiar role that strategies play in the way that people reason about analogous planning cases.

During and after the completion of this large-scale representation work, we made several observations that contribute to our theoretical understanding of strategies. In particular, two points are presented here that are specifically targeted at the cognitive science research area of analogical reasoning. First, we argue that strategies are themselves portions of the relational structure that serves as the basis for analogical reasoning about planning cases. Second, mental representations of strategies include references to the reasoning processes of intentional agents, providing us with a means of describing the models that people have of their own reasoning processes and those of others. Both of these arguments are developed in the following two sections.

Strategies are Relational Structures Shared by Analogous Planning Cases

In June of 1941, Germany invaded Ukraine with three million troops, threatening to advance eastward into Russia. Soon after, Soviet leader Joseph Stalin announced a scorched-earth policy for Ukraine, ordering that retreating Ukrainians destroy everything that could be of value to the advancing German army, including grains, fuel, and engines that could not be transported east to Russia. In an analogous case, Iraq invaded their oil-rich neighbor Kuwait in August of 1990, leading to the Persian Gulf war. The following January, the United States launches an attack against Iraq, and Saddam Hussein responded by blowing up Kuwaiti oil wells and dumping millions of gallons of oil into the Persian Gulf.

Analogies of exactly this sort have been the subject of a number of experimental studies of analogical reasoning (especially Spellman & Holyoak, 1992), and competing theories have been proposed as cognitive models for this sort of mental processing. The two theories that have received the most attention are Structure-mapping theory (Gentner, 1983, 1989) and Multiconstraint theory (Holyoak & Thagard, 1989, 1995). Although they have their differences, the two theories agree that analogical reasoning is based on structural similarity, the similarity of the systems of relationships that exist between the represented entities in two different cases. Both agree on the constraint of *one-to-one correspondence* between represented entities in analogous cases, e.g. Kuwait could potentially correspond to Germany, Ukraine, or Russia, but not more than one of these in any given analogical mapping. Both also agree on the constraint of the *systematicity*, requiring that sets of relationship correspondences have consistent argument correspondences, e.g. because Iraq is an argument in both the relationships of invading and destroying, both of these relationships cannot be a part of the same system of analogical mappings to the WWII case, where Germany did the invading and the Ukraine did the destroying.

Given the constraints of one-to-one correspondence and systematicity, these theories predict that the strength of any given analogy is strongly dependent on the way that the cases are represented. If we assume representations that are too sparse, we risk predicting that the analogy between a Persian Gulf war case and a WWII case would be a relatively weak one. If Germany is mapped to Iraq, then we have correspondence between the relationship of invading (Germany/Ukraine and Iraq/Kuwait) and the relationship of contained-within (Ukraine/Grains and Kuwait/oil wells). If the Soviet Union is instead mapped to Iraq, then we have a correspondence between the relationship of destroying (Soviet Union/grains and Iraq/oil wells) and that of possession or ownership (Soviet Union/Grains and Iraq/oil wells).

However, this analogy is intuitively very strong and unambiguous. The obvious mapping is that Iraq is like the Soviet Union, as their decision to destroy the Kuwaiti oil wells was analogous to when Stalin ordered the destruction of resources in the Ukraine. These cases are two examples of the *exact same strategy* - instances of an abstract pattern of planning behavior that is so prevalent in our culture that we've given it a name, *scorched earth policy,* so that we could refer to it again and again in analogous cases, whether they appear in warfare or in completely different domains such as politics or business. To account for the comparative strength over this interpretation of the analogy over others, we must assume that the representations of these cases are much richer.

When we consider the abstract similarities that are found in the planning that is done in these two cases, the structural alignment becomes clear. The agent that is doing the planning in these cases (Stalin/Hussein) has some adversarial relationship with some other agent (Hitler/Bush). This planning agent imagines a likely future where the adversary acquires possession of some resources (grain/oil) that are currently possessed by the planner. They imagine that the adversary will make use of these resources to further the pursuit of their adversarial plan (march on to Russia/control the middle east). They decide that the best plan is to do something (destroy grain/blow up oil wells) that will cause these resources to be destroyed, or to make it impossible that the adversary could make use of them, and to do so before the adversary gains possession.

While the rich relational alignment between these two military examples is described using natural language in the preceding paragraph, a corresponding mental representation language would necessarily include structures to refer to the adversarial relationship, the imagination of a likely future, the acquisition of resources, the expenditure of resources in an adversarial plan, the goal of disabling a resource, and the execution deadline. It is this collection of relationships that constitutes the representation of the strategy, and which also makes a significant contribution to judgments of analogical similarity to *every other case* that describes an instance of this sort of strategic behavior.

To clarify, we would like to point out that not all of the relational structure shared between analogous planning cases can be thought of as part of a strategy. Certainly there are analogous planning cases that are so not because of any similarity in the strategies of the participating agents. For example, a case where a person marries someone just before the other person

wins the lottery may be analogous to a corporation that acquires another business just before it has an unexpected licensing windfall, but the commonalties in these cases have more to do with unforeseen benefits than strategic thinking. In contrast, if it turns out that the person and the parent corporation both had selected their candidate acquisitions based on a perception of how lucky the candidates were, we would say that they shared the same strategy - in reference to the portion of the shared relational structure that concerned the planning processes of these agents.

The research opportunity that is evident here concerns the apparent ease that people have in making casual references to large portions of shared relational structure, removing these portions from their context to be considered independently, and assigning to them names like *scorched earth policy* when they are particularly interesting for one reason or another. This ease with strategies enables researchers to collect whole catalogs of naturally occurring analogical mappings, and to argue about how they could be represented on a much larger scale than was possible in previous knowledge representation debates.

Strategy Representations Include the Reasoning Processes of Agents

The formal case representations that have appeared in computer models of analogical reason consist almost entirely of propositions about the external world (Falkenhainer *et al*., 1989; Forbus *et al*., 1994; Holyoak & Thagard, 1989; Holyoak *et al*., 1994). The most compelling examples that support their corresponding theories of analogical reasoning are often in the domain of physical systems, where existing representational theories support the way that these example cases were represented, notably Qualitative Process Theory (Forbus, 1984). Likewise, the case representations in the computer models of analogical reasoning that seem the most contrived are those that are more story-like in nature, involving the intentional behavior of intelligent agents.

A somewhat notorious example of this problem can be seen in the "Karla the hawk" stories that were used in support of the Structure-Mapping Engine (Falkenhainer *et al*., 1989). The main story concerned the actions of Karla, an old hawk, and her encounter with a hunter who had attempted to shoot Karla with an arrow, but had missed because the hunter's crude arrow lacked the feathers needed to fly straight. Karla knew that the hunter wanted her feathers so she offered some of hers to him, and he gratefully pledged never to shoot at a hawk again. In the dozen propositions that were authored to represent this story, emphasis was placed on describing the actions that where done (e.g. seeing, attacking, offering, obtaining, promising) with only a single, highly ambiguous predicate of *realizing* to refer to the reasoning processes that Karla undertook.

What is lacking from these representations is the richness of planning and reasoning that is ascribed to these characters when we read stories like this - the conceptual glue that allows us to make sense of the story in the first place. Much of this knowledge can be packaged into the form of a single strategy, the one that we guess that Karla had in mind when she offered the feathers. Figure 2 offers a preformal representation of this strategy, where the capitalized and italicized words and phrases come from the set used in our large-scale strategy representation project. Of course, it is possible to imagine that Karla wasn't thinking strategically at all - often characters in this genre of fable-like stories seem to stumble across some course of action by mere chance, without thinking things through. However, it is difficult to imagine that a reader could be as ignorant. Indeed, it is the planning lessons that can be learned from stories of this genre that make them compelling and valuable.

What should be noticed in Figure 2 is that the strategy

Karla's Strategy: Turn enemies into friends by improving their capabilities
The planner has an *Adversarial Relationship* with another agent that has had an *Execution failure* of an *Adversarial plan*, with a *Cause of failure* of *Instrument failure*. The planner *Envisions* that this agent will *Attempt the execution* of this *Adversarial plan* in *Future states* against the planner and *Envisions a possibility* that this plan will be *Successfully executed.* The planner has a *Partial plan* to *Reduce the probability* of *Instrument failure* in *Future executions* of the *Adversarial plan* by this other agent that includes *Modifying the instrument* and includes the *Instrumental use* of *Resources* of the planner. The planner executes a plan to *Make an offer* to the agent where the *Offered action* is that the planner *Transfers possession* of *Resources* that *Enable* the agent to execute the *Partial plan*. The planner then *Adds the expectation* that *If it is the case* that the agent *Accepts the offer*, that the agent will *Terminate their role* in the *Adversarial relationship*, which will cause them to *Abandon plans* that are *Adversarial plans* against the planner. The planner then *Adds a threat* that the *Expectation* is a *False expectation*, and that the agent will execute the *Partial plan* followed by the execution of the *Adversarial plan*.

Figure 2. The strategy of Karla the hawk as a preformal representation

contains a significant amount of references to mental states of both the planner and the adversary. There is the imagining of future states, of a partial plan not entirely worked out, an expectation of the consequences of an action, and an explicit threat of what might happen if this expectation is wrong. Each of these mental states is critical to the understanding of the story, and we argue that they should be included in the representation of this case to explain analogies to cases where only the strategy is shared.

Just as analogies in physical systems are based on mental models of processes in physical domains, analogies in intentional domains include assertions that only make sense with respect to a mental model of agents and intentional behavior. In arguing for the inclusion of these sorts of assertions in case representations in intentional domains, we are also making the argument that people have a rich model of agents and their reasoning processes.

A great deal of attention has been directed toward developing models of agents and intentional behavior among decision theorists and artificial intelligence logicians, often centered around the notion of Belief, Desire and Intention (BDI) agents. Formalizations of these models (e.g. Cohen and Levesque, 1990) typically strive to maximize the number of simplifying assumptions in order to retain the ability to prove related theorems, but to do so without sacrificing the expressiveness required to compute useful functions. The engineering value of this approach is demonstrated when theories lead to practical applications (see Rao & Georgeff, 1995), but we caution against this approach for the purpose of cognitive modeling. Certainly there are times where the mental representations that people have may appear to exhibit the qualities of elegance and simplicity, but our aim should be to describe the mental models of people as they are - without simplifying assumptions - if we are to understand and predict how they are manipulated by cognitive processes.

In our own investigation of mental representation through the lens of strategies, we have found that the models that people have of other agents and their reasoning processes is enormously complex in comparison to previous formalizations. From the most generalized perspective, the model appears to be comprised of many of the components that are commonly proposed. These components include representations of the current state, agents and their goals, the plans that are held by these agents, the envisionments that these agents construct of past and future states, of the plan construction process, the making of decisions, the scheduling of behaviors, the monitoring of events, and the process of executing an intended plan in the world. Given a closer look, we find that each of these model components is extremely complex. For example, the representations that people have of the process of executing an intended plan are rich enough to refer to the sensations that agents experience during an execution, and to remark on whether it had a natural quality to it or not - as in reference to a concert pianist that finds a particular passage in a piece cumbersome due to an awkward fingering that they selected. Indeed, a strategy that concert pianists employ is to reduce the risk of performance blunders by explicitly identify the sections of a musical piece that give rise to these sensations of awkwardness, and to rework their plan of execution for these sections so that they have a more natural quality.

Conclusions: Analogy as a Tool

"The association of solved problems with those unsolved may throw new light on our difficulties by suggesting new ideas. It is easy to find a superficial analogy which really expresses nothing. But to discover some essential common features, hidden beneath a surface of external differences, to form, on this basis, a new successful theory, is important creative work." (Einstein & Infeld, 1938)

Einstein and Infeld presented this idea to justify making an analogy between a violin string and the wavelike properties of elementary particles. The essential common features that Einstein and Infeld discovered were the concepts of a *standing wave* and a *moving wave*, and the new successful theory that they were forming was that of probability waves in quantum physics.

Einstein and Infeld's quote reveals something about the way that they approached scientific problem solving, but from a cognitive science perspective, we might also view it as a proposed methodology for forming theories of people's mental models. That is, by analyzing analogous cases, discovering the essential common features that are hidden beneath a surface of external differences, we can form theories of the mental models that people have that cause these cases to be analogous.

In this paper we have argued that planning strategies are particularly appropriate as the subject of analysis using the tools of analogy. Strategic analogies reveal features of our mental models of human planning, and in doing so, challenge our cognitive theories and models of intelligent planning and problem solving behavior. It is this hope that led us to author three hundred and seventy two representations of strategies in ten different planning domains, where each representation attempted to define the features of the planning situation that were common among all analogous instances of the strategy.

Einstein and Infeld made reference to two important activities in the quoted text. The first, to discover

essential common features of analogous cases, has been accomplished on a large scale for instances of strategies. Achieving the second, to form a new successful theory, will require a substantial amount of additional work. This paper has argued for the inclusion of two claims in the future successful theories that are developed. First, we argued that strategies are themselves portions of the relational structure that is the basis of similarity between analogous planning cases. Second, the representations of these strategies include propositions about the reasoning processes of the agents employing them, giving researchers a investigative lens to examine the rich mental models that people have of others and of themselves.

References

Cohen, P. & Levesque, H. (1990) Intention is choice with commitment. *Artificial Intelligence*, 42(3).

Collins, G. (1986) Plan creation: Using strategies as blueprints. Ph.D. dissertation, Yale University.

Einstein, A. & Infeld, L. (1938) The evolution of physics from early concepts to relativity and quanta. New York: Simon & Schuster, p. 273.

Falkenhainer, B., Forbus, K., & Gentner, D. (1989). The Structure-Mapping Engine: Algorithm and Examples. *Artificial Intelligence*, 41, pp 1-63.

Forbus, K. D., Ferguson, R. W., & Gentner, D. (1994). Incremental structure-mapping. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, 313-318. Hillsdale, NJ: Lawrence Erlbaum Associates.

Forbus, K. (1984). Qualitative process theory. *Artificial Intelligence*, 24, pp 85-168.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155-170.

Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*, pp. 199-241. London: Cambridge University Press.

Gil, Y. & Blythe, J. (2000) PLANET: A sharable and reusable ontology for representing plans. *AAAI 2000 workshop on Representational issues for real-world planning systems*.

Gordon, A. (2001) The representational requirements of strategic planning. *Proceedings of the Fifth Symposium on Logical Formalizations of Commonsense Reasoning*.

Holyoak, K. & Thagard, P. (1995) *Mental leaps: Analogy in creative thought*. Cambridge, MA: MIT Press.

Holyoak, K. & Thagard, P. (1989) Analogical mapping by constraint satisfaction. *Cognitive Science*, 13, 295-355.

Holyoak, K., Novick, L, & Melz, E. (1994) Component processes in analogical transfer: Mapping, pattern completion, and adaptation. In K. Holyoak & J. Barnden (Eds.), *Advances in connectionist and neural computation theory, Vol. 2: Analogical connections*. (pp. 113-180). Norwood, NJ: Ablex.

Jones, E. (1992) The flexible use of abstract knowledge in planning. Ph.D. dissertation. Yale University.

McDermott, D. (2000) The 1998 AI Planning Systems Competition. *AI Magazine*, 21(2), 35-55.

Owens, C. (1990) Indexing and retrieving abstract planning knowledge. Ph.D. dissertation. Yale University.

Roa, A. & Georgeff, M. (1995) BDI agents: From theory to practice. *Proceedings of the First International Conference on Multiagent Systems*.

Spellman, B. & Holyoak, K. (1992) If Saddam is Hitler then who is George Bush? Analogical mapping between systems of social roles. *Journal of Personality and Social Psychology*, 62, 913-933.

Tate, A. (1998) Roots of SPAR - Shared planning and activity representation. *Knowledge Engineering Review*, 13, 121-128.

Superstitious Perceptions

Frédéric Gosselin (GOSSELIF@PSY.GLA.AC.UK)
Philippe G. Schyns (PHILIPPE@PSY.GLA.AC.UK)
Lizann Bonnar (LIZANN@PSY.GLA.AC.UK)
Liza Paul (LIZA@PSY.GLA.AC.UK)
Department of Psychology, University of Glasgow, 58 Hillhead Street
Glasgow, Scotland, G12 8QB UK

Abstract

It has long been observed that we sometime perceive complex scenes in blots, rocks, or clouds, but the phenomenon has attracted little scientific attention. We propose that a weak–or superstitious–match between a memory template and a sparse stimulus is responsible for such perceptions. We provide reverse-correlation evidence for this theory.

Introduction

> If you look at walls that are stained or made of different kinds of stones […] you can think you see in them certain picturesque views of mountains, rivers, rocks, trees, plains, broad valleys, and hills of different shapes […] battles and rapidly moving figures, strange faces and costumes, as well as an infinite number of things […]
>
> (Leonardo da Vinci, *Notebooks*)

We have all seen a human face or a landscape in a cloud floating by, in a pebble lying on a beach, or in blots on a wall. Notorious examples of this phenomenon include the Mars channels and the Man on the Moon; Hermann Rorschach has even made it the basis of a projective test. The earliest known reference to the phenomenon reaches back as far as classical antiquity, and thousands of others have been enumerated (Janson, 1973; Gombrich, 1960). Given this human fascination for the phenomenon, it is surprising how little–if any–scientific attention it has received. Here, we provide evidence that these perceptions result from a weak–or superstitious–match between a memory template and a sparse stimulus. Beyond the anecdotes, a rigorous study of superstitious perceptions could reveal important properties of internal object representations. It is one aim of our research to illustrate this point.

We instructed naïve observers to decide whether one particular target (the letter 'S' in Experiment 1 and a smiling face in Experiment 2) was present or not in stimuli. No signal was ever presented in the stimuli. Each stimulus comprised only two-dimensional static bit "white" noise. White noise has several desirable properties: It has equal energy across the entire spatial frequency spectrum and does not correlate across trials. In other words, white noise does not in itself represent coherent structures in the image plane and across trials.

These properties make white noise the perfect basis for reverse correlation (see Appendix), a statistical technique that uses noise to derive the information the observer *uses to respond* in a particular visual task (e.g., Ahumada & Lovell, 1971; Beard & Ahumada, 1998; Neri, Parker & Blakemore, 1999; Gold, Murray, Bennett & Sekuler, 2000). In Experiment 1, we used reverse correlation (supplemented with careful debriefing) to assess the properties of the letter 'S' that the observers superstitiously perceived (remember that they only saw white noise). Experiment 2 replicated the findings in the more realistic case of faces.

Experiment 1: 'S' as in Superstitious

In this experiment, we asked a first subject to detect in white noise the presence of a black 'S' on a white background filling the image. As just explained, only bit noise was presented.

Method

Subject

One 24-year old female student from the University of Glasgow with normal vision was paid £50 to participate in this study. She was an experienced psychophysical observer, but had no knowledge about the goals of the experiment.

Procedure

The experiment ran on a Power PC Macintosh using a program written with the Psychophysics Toolbox for Matlab (Brainard, 1997; Pelli, 1997). It comprised 20,000 trials equally divided into 40 blocks. The subject took two weeks to complete the experiment. A trial consisted in the presentation of one 50 x 50 pixels (2 x 2 deg of visual angle) static bit noise image with a black-pixel density of 50%. No signal was ever presented. The subject was told, however, that she was participating in a detection experiment. She was instructed to say whether or not a black letter 'S' on a white background filling the image was present. No more detail was provided about the 'S'. We told her that 50% of the trials were positive. The subject was under no time pressure to respond.

When the 20,000 trials were completed, we debriefed the subject. We asked her the following questions: How often did she see the letter? When she

saw it, how noisy was it? What strategy did she use to respond?

Results and discussion

On 22.7% of the trials the subject pressed on the 'yes' key, indicating that an 'S' was present. During debriefing, she said that she saw an 'S' each time she responded positively, and she estimated the quantity of noise in the letter 'S' to vary between 30% and 50%. She summarized her strategy as follows: "I simply waited to see if the S "jumped out at me"."

All the static bit noise images leading to a 'yes' response were added together and so were those leading to a 'no' response. The two resulting images, the 'yes' and the 'no' images, were normalized. A raw classification image was then computed by subtracting the normalized 'no' image from the normalized 'yes' image. This classification image is the linear template that best explains the behavior of the subject in the least square sense of the term (see Appendix).

There is an objective method to understand the information that drove the illusory perceptions of the 'S' in the experiment. As explained earlier, white noise is completely unbiased. If the observer responded randomly (i.e., without having the illusion of the presence of an 'S'), the classification image would itself be unbiased. From this reasoning, any bias appearing in the spectral analysis of the observer's classification image should indicate the structures underlying the illusory perceptions. The spectral analysis reveal a bias for information concentrated between 1 and 3 cycles per image, with a peak at 2 cycles per image (see arrow in Figure 1a). This is consistent with Solomon and Pelli's (1994) finding that letter identification is most efficient around 3 cycles per letter.

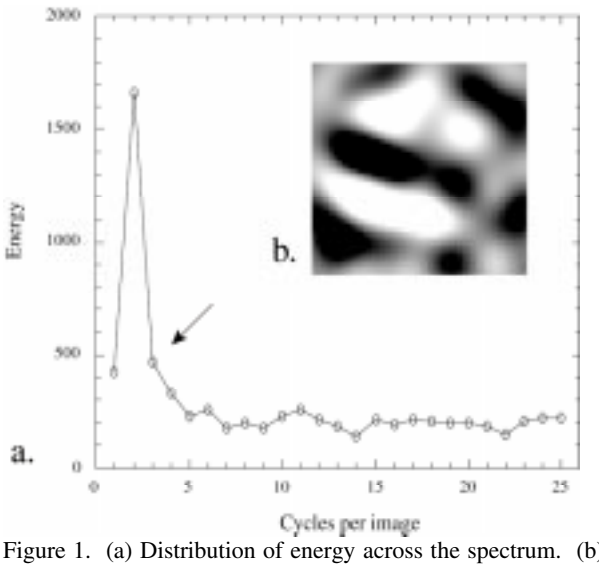


Figure 1. (a) Distribution of energy across the spectrum. (b) Classification image low-passed at 3 cycles per image.

We can visualize the information that drove the illusory detection by filtering the raw classification image with a low-pass Butterworth filter with a cutoff at 3 cycles per image. To provide a better depiction, we further remove all the outlier pixel intensities (two standard deviations away from the mean). The resulting image is a black 'S' on a white background.

To summarize, we have induced illusory perceptions of an 'S' by asking one subject to detect this letter in noise. Unknown to her, the stimuli did not comprise the letter, but only white noise. If the subject had been performing only according to the stimulus (i.e., in a bottom-up manner), her classification image should have had the same properties as noise—i.e., having identical energy across all spatial frequencies. However, there was a marked peak of energy between 1 and 3 cycles per degree that could only arise from top-down influences on the interpretation of white noise. Further analyses revealed the precise shape of the letter that the subject thought she saw. Specifically, it is worth pointing out that the best depiction of the information used

Experiment 2: Simile smile

In Experiment 2, we generalized the technique to a more complicated stimulus, using another subject. The task was to discriminate between a smiling and non-smiling face. However, the face presented in noise had no mouth whatsoever.

Method

Subject

One 26-year old female student at the University of Glasgow with normal vision was paid £50 to take part in this study. She was naïve with respect to the goals of the experiment, but was an experienced psychophysics observer.

Procedure

The experiment ran on a Macintosh G4 using a program written with the Psychophysics Toolbox for Matlab (Brainard, 1997; Pelli, 1997). It consisted in 20,000 trials equally divided in 40 blocks. The subject took three weeks to complete the experiment. In each trial, one sparse image spanning 256 x 256 pixels (5.72 x 5.72 deg of visual angle) was presented. This image comprised 27.5% of the black pixels of the contours of a mouthless face (see the white marker in Figure 2b) randomly sampled and, for the remainder, of bit noise with the same density of black pixels. No signal was therefore presented in the mouth area.

The subject was instructed to decide whether the face was smiling or not–no detail was provided regarding the alternative expressions. This ensured that the subject focused on seeking information for "smile". We also told her that the face would be smiling in 50%

of the trials. The subject was under no time pressure to respond. Following the 20,000 trials, we debriefed the subject as in Experiment 1.

Results and discussion

On 7.07% of the trials the subject pressed on the 'yes' key, indicating that the "noisy" face was smiling. During debriefing, she explained that she had been very conservative and that she had only responded 'yes' when she was absolutely certain that the face was indeed smiling. The subject looked for teeth and used the eyes and the nose to localize the mouth.

All the static bit noise images leading to a 'yes' response were added to form a 'yes' image, and all those leading to a 'no' were added to form a 'no' image. A raw classification image was then computed by subtracting the normalized 'no' image from the normalized 'yes' image.

The distribution of energy in the spectrum for the raw classification image is represented in Figure 2a. The energy is concentrated in the bandwidth ranging from 1 to 20 cycles per image (from 0.35 to 12.29 cycles per face–see arrow in Figure 2a). This roughly corresponds to the most efficient bandwidth found by Bayer, Schwartz and Pelli (1998) in an expression identification task (i.e., maximum efficiency centered at 8 cycles per face).

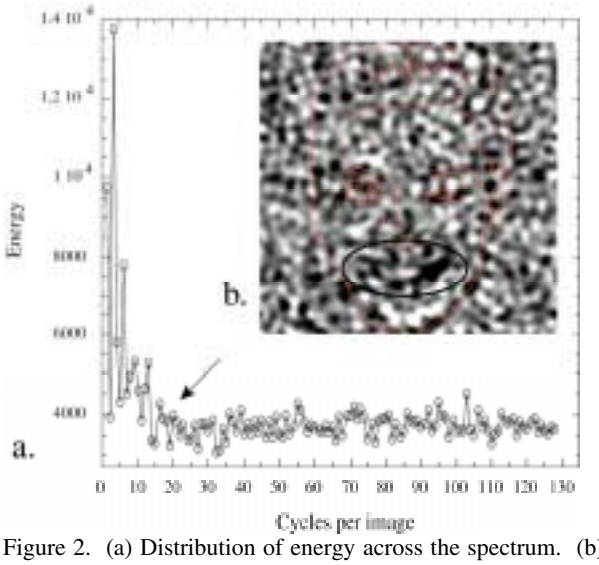


Figure 2. (a) Distribution of energy across the spectrum. (b) Classification image low-passed at 20 cycles per image.

Figure 2b is the raw classification image low-passed at 20 cycles per image with a Butterworth filter–with outlier pixel values removed, followed by a normalization. A white mouthless face marker has been superimposed on filtered classification image. A smile revealing teeth is clearly visible (see circled area in Figure 2b).

Conclusion

The evidence we have gathered in two experiments corroborates the idea that superstitious perceptions result from a weak match between a memory template and a sparse stimulus. We have shown that we could induce superstitious perceptions of a letter ('S', Experiment 1) and part of a face (a mouth expressing a smile, Experiment 2) in bit noise. Reverse correlation demonstrated that observers in these experiments used information from memory ressembling an 'S' and a smile, respectively. It is important to stress that this information did not originate from the signal, by from their memory. It is only because these memory representations are partially correlated with white noise that the superstitious perceptions occur. But then, because white noise is weakly correlated with every visual stimulus, this technique could in principle extend to depicting a wide range of visual representations. In our experiments, these representations had properties expected from what is know in the recognition literature. So, the superstitious perceptions were not random hallucinations, but instead well-constrained perceptions derived from specific knowledge.

Superstitious perceptions could therefore be used to explore the properties of representations in the absence of any bottom-up information.

Acknowledgements

This research was partially funded by ESRC grant R000 237 901.

References

Ahumada, A. J. & Lovell, J. (1971). Stimulus features in signal detection. *Journal of the Acoustical Society of America*, **49**, 1751-1756.

Bayer, H. M., Schwartz, O. & Pelli, D. (1998). Recognizing facial expressions efficiently. *IOVS*, **39**, S172.

Beard, B. L. & Ahumada, A. J. (1998). "A technique to extract the relevant features for visual tasks" In B. E. Rogowitz and T. N. Pappas (Eds.), Human Vision and Electronic Imaging III, SPIE Proceedings, **3299**, 79-85.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10, 433-436.

Gold, J., Murray, R. F., Bennett, P. J. & Sekuler, A. B. (2000). Deriving behavioral receptive fields for visually completed contours. *Current Biology*, **10** (11), 663-666.

Gombrich, E. H. (1960). *Art and Illusion: A Study in the Psychology of Pictorial Representation.* London: Phaidon Press Ltd.

Janson, H. W. (1973). "Chance images" In Philip P. Wiener (Ed.), *Dictionary of the History of Ideas.* New York: Charles Scribner's Sons.

Neri, P., Parker, A. J. & Blakemore, C. (1999). Probing the human stereoscopic system with reverse correlation. *Nature*, **401** (6754), 695-698.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, **10**, 437-442.

Solomon, J. A. & Pelli, D. G. (1994). The visual filter mediating lettre identification. *Nature*, **369**, 395-397.

Sprent, P. (1969). *Models in Regression and Related Topics.* London: Methuen & Co Ltd.

Appendix

We suppose that the observer matches two vectors at each trial of the experiment: a stimulus vector of dimensionality k and a template vector $\mathbf{ß}$ of the same dimensionality representing the memorized pattern to match against the input (e.g., the letter 'S' or a smiling face).

Suppose further that we arrange the n stimuli of the experiment in the $n * k$ matrix $\mathbf{X}$. The behavior of the observer for the whole experiment is described by

$$\mathbf{y} = \mathbf{ß}\mathbf{X}+\varepsilon,$$

where $\mathbf{y}$ is an n-dimensional vector of decision responses, and ε is an n-dimensional vector of "error" random variables with $E(\varepsilon) = \mathbf{0}$ and $V(\varepsilon) = \sigma^2\mathbf{I}$.

For simplicity, the "target present" and "target absent" responses in $\mathbf{y}$ as well as the white and black pixels in $\mathbf{X}$ are encoded with values of 1 and –1, respectively.

Given that we know $\mathbf{X}$ and can observe $\mathbf{y}$, we can resolve the linear system of equations by finding $\mathbf{ß}$. The *least square* solution requires that we minimize the scalar sum of squares

$$\mathbf{S} = (\mathbf{y}-\mathbf{X}\mathbf{ß})'(\mathbf{y}-\mathbf{X}\mathbf{ß})$$

for variations in $\mathbf{ß}$. Differentiating, we have

$$2\mathbf{X}'(\mathbf{y}-\mathbf{X}\mathbf{ß}) = \mathbf{0},$$

which gives, for our least square estimator, the vector

$$\mathbf{ß} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

This is the logic of standard multiple regression (e.g., Sprent, 1969). Because our stimulus vectors are uncorrelated, we have

$$(\mathbf{X}'\mathbf{X})^{-1} = (k\mathbf{I})^{-1} = k^{-1}\mathbf{I},$$

Therefore,

$$\mathbf{ß} = k^{-1}\mathbf{X}'\mathbf{y}.$$

Leaving the constant k aside, this equation reduces to summing all stimuli that led to a 'yes' response and subtracting from it the sum of all the stimuli that led to a 'no' responses. This is the essence of reverse correlation.

Words and Shape Similarity Guide 13-month-olds' Inferences about Nonobvious Object Properties

Susan A. Graham (grahams@ucalgary.ca)
Department of Psychology, University of Calgary
Calgary, AB T2N 1N4 Canada

Cari S. Kilbreath (cskilbre@ucalgary.ca)
Department of Psychology, University of Calgary
Calgary, AB T2N 1N4 Canada

Andrea N. Welder (anwelder@ucalgary.ca)
Department of Psychology, University of Calgary
Calgary, AB T2N 1N4 Canada

Abstract

We examined the influence of shape similarity and object labels on 13-month-old infants' inductive inferences. In two experiments, infants were presented with novel target objects with or without a nonobvious property, followed by test objects that varied in shape similarity to the target. When objects were not labeled, infants generalized the nonobvious property to test objects that were highly similar in shape (Expt. 1). When objects were labeled with novel nouns, infants generalized the nonobvious property to both high shape similarity and low shape similarity test objects (Expt. 2). These findings indicate that infants as young as 13 months of age expect those objects which share the same shape or the same label to possess the same nonobvious property.

Introduction

Inductive reasoning involves invoking the premise that things that are true for one member of a category (e.g., the blue ball bounces) will hold true for other members of the same category (e.g., therefore all balls bounce; Moore & Parker, 1989). The ability to reason inductively is an invaluable cognitive skill, allowing an individual to generalize knowledge to new instances and new situations. In recent years, a great deal of empirical attention has been devoted to examining preschoolers' inductive reasoning abilities, with particular focus on the nature of the categories that guide their inferences. In a typical inductive generalization task, preschoolers are taught a fact about a target object and then are asked whether that fact can be generalized to other test objects. Using this methodology, studies have demonstrated that by 2-1/2-years of age, children can reason inductively about object properties in remarkably sophisticated ways (e.g., Gelman, 1988; Gelman & Markman, 1986, 1987; Kalish & Gelman, 1992). For example, Gelman and Coley (1990) found that 2-1/2-year-olds will overlook perceptual similarity and generalize properties on the basis of shared underlying kind when the target and test objects are given the same count noun label.

In recent years, researchers have begun to examine the development of inductive capabilities during the infancy period using the generalized imitation paradigm. In a typical task, an experimenter will first model a specific action on a target object. He or she will then hand infants test objects and observe whether or not they imitate the target action on the various objects. Studies using this paradigm indicate that infants as young as 9 months of age will draw inferences about nonobvious object properties based on knowledge gained during the experimental session (Baldwin, Markman, & Melartin, 1993). Furthermore, research suggests that both perceptual similarity and conceptual knowledge may play a role in guiding infants' inferences (Baldwin et al., 1993; Mandler & McDonough, 1996, 1998). In a recent series of studies (Welder & Graham, in press), we found that 18-month-old infants will rely on shared shape similarity to guide their inductive inferences about novel objects' nonobvious sound properties when no other information about category membership is available. More importantly, however, when infants were provided with information about conceptual category membership in the form of shared object labels, shape similarity was either attenuated in significance (in the case of novel labels) or disregarded (in the case of familiar labels). These findings indicate that 18-month-old infants can make inductive inferences about object properties based on a conceptual notion of object kind. Furthermore, these findings suggest that infants as young as 18 months of age recognize the conceptual information conveyed by object labels. That is, they recognize that noun labels supply information about underlying object kind and that members of the same kind share nonobvious properties.

In the present studies, we pursued the investigation of infants' inductive abilities, with specific focus on the reasoning abilities of infants who are just beginning to acquire productive language. First, we examined whether 13-month-olds, like 18-month-olds, will rely on shared shape similarity to generalize nonobvious object properties, in the absence of other information about

object kind (Experiment1). Second, we examined whether 13-month-olds will rely on shared object labels to direct their inductive inferences (Experiment 2). In particular, we examined whether infants will extend a nonobvious property on the basis of a shared object label, even if the objects differ in shape.

In both experiments, we employed a generalized imitation paradigm to examine infants' inductive abilities (see also Baldwin et al., 1993; Mandler & McDonough, 1996, 1998). We presented infants with novel target objects that possessed nonobvious properties (e.g., a cloth-covered object that squeaked when squeezed). The experimenter demonstrated the nonobvious property using a specific target action and then presented infants with test objects which varied in their degree of perceptual similarity to the target. We reasoned that if infants considered test objects to be members of the same category as the target, they would expect the test objects to share the same nonobvious property as the target. That is, infants' imitation of a target action on test objects would provide evidence of inductive reasoning.

Experiment 1

The goal of Experiment 1 was to examine the role of shape similarity in guiding infants' generalization of nonobvious object properties when they are presented with novel object categories. Infants were presented with object sets consisting of a target object followed by a high similarity match, a low similarity match, and a dissimilar object in three within-subjects conditions. The high and low similarity matches within each set varied in shape and color but shared similar textures. The dissimilar objects in each set, however, differed from the target object in texture, shape, and color. The dissimilar objects were included to ensure that infants' inductive generalizations were specific to objects that they perceived as belonging to the same category and to ensure that infants were not merely imitating the experimenter's actions on any object, regardless of whether an expectation was generated.

We presented infants with target and test objects in three within-subjects expectation conditions. In the surprised condition, the target object possessed an interesting sound property (e.g., squeaked when squeezed), but the test objects were disabled so that they could not exhibit the property (e.g., could not squeak when squeezed). This condition was of particular interest, as infants' performance would indicate whether they expected test objects to possess the same nonobvious property as the target. In the baseline condition, neither the target nor the test objects possessed the interesting property (e.g., neither produced a squeak sound). This condition provided a baseline measure of infants' exploratory actions. A comparison of infants' performance in the baseline condition to the surprised condition would indicate whether the target property of the target object was, in fact, nonobvious upon visual inspection. In the predicted condition, both the target and test object possessed the property (e.g., both could squeak). This condition was included to preclude the development of the expectation that all test objects were disabled (as all test objects in both the surprised and baseline conditions were), leading infants to become bored or frustrated with the stimuli.

We expected that infants would use shape similarity to guide their generalizations of the nonobvious property as object shape is easily perceived and is often an excellent index of object kind (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). More specifically, we predicted that the greater the degree of shape similarity between a test object and a target object, the higher the frequency of target actions performed on that test object.

Method

Participants Participants were 20 infants ranging in age from 12.07 months to 13.92 months (M = 12.73; SD = .53). Ten infants were male and 10 were female.

Stimuli Four objects were used for the warm-up trials: a garlic press, a roller ball, a clicking clock, and a clothesline pulley. Three object sets (a "squeaking" set, a "ringing" set, and a "rattling" set) were created for use in the imitation task. There were four objects in each set: a target object, and three test objects (a high similarity object, a low similarity object, and a dissimilar object). The high similarity test objects possessed the same shape and texture as the target object but differed in color. The low similarity objects shared the same texture as the target object but differed in shape and color. The dissimilar object shared no properties in common with the target object. The dissimilar objects included a plastic orange file (squeaking set), a small white strainer (ringing set), and a plastic green hose splitter (rattling set). There were two versions of each target and test object in each set: a functional version that could produce the target sound and a nonfunctional version that was disabled and thus unable to produce the target sound. In the functional squeaking object set, objects could produce a squeaking noise when squeezed. In the functional ringing set, objects could produce a ringing noise when tapped. In the functional rattling set, objects could produce a rattling noise when shaken.

To establish whether test objects could reliably be categorized as high or low in shape similarity relative to the target object, 15 adults rated the similarity of each test object to its target. The adult ratings followed the expected pattern. That is, the high and low similarity test objects in each object set were perceived as significantly different in shape from one another (all t-tests: $p < .05$), in the direction intended.

Design For each infant, one of the three object sets was presented in the surprised condition, one set was presented in the baseline condition, and one set was presented in the predicted condition. The specific object set assigned to the surprised, baseline, and predicted conditions was counterbalanced across infants.

The imitation task was comprised of three blocks of three trials each: one trial in the surprised condition, one trial in the predicted condition, and one trial in the baseline condition. Each object set was presented once within each block. That is, one of the test objects (e.g., a high similarity object) from a given set was presented in the first block, another (e.g., a low similarity object) was presented in the second block, and a third (e.g., a dissimilar object) was presented in the third block. The order of presentation of test objects was randomized within each block and order of presentation of expectation condition within each trial block was counterbalanced across infants.

Procedure Infants were seated in their parent's lap at a table in a testing room with the experimenter seated across from them. Before testing began, the experimenter instructed parents to interact with their infant as little as possible and not to direct their infant's attention to the objects. Parents were also instructed to silently place objects back on the table within the infant's reach if objects were dropped on the floor near the parent or if the infant handed objects to the parent. All sessions were videotaped for coding purposes.

During the warm-up phase, the experimenter demonstrated a target property of each of the warm-up objects to the infant and asked the parent to do the same. After demonstrating the target property, parents silently handed the object to their child for him/her to imitate the actions observed.

During the test phase, the experimenter began each trial by presenting infants with one of the target objects from a given object set. She introduced the object (e.g., "Look at this one!") and demonstrated the nonobvious property of the target object five times (e.g., shaking the rattle). Only the properties of target objects in the surprised and predicted conditions were demonstrated (as the target objects in the baseline condition did not possess the property). The experimenter handed the object to the infant's parent who demonstrated the property of the target object twice. The parent then passed the object to the infant. After a period of 10 seconds, the experimenter retrieved the object and placed it within the infant's view, but out of reach. The experimenter then presented the infant with a test object and infants were allowed to explore the test object for 20 seconds. This same procedure was repeated for each of the other 8 trials. The target object from each object set was reintroduced to infants on each trial; however, parents only demonstrated the property the first time a target object was introduced. The experimenter continued to demonstrate the target object's property on each trial. If an object was dropped off the table or passed/thrown out of the infants' reach during the session, the experimenter quickly placed the object back within their reach. Time lost due to these actions was not compensated for, as they were considered to be intentional actions of frustration or disinterest (see Oakes, Madole, & Cohen, 1991).

Coding Coders, blind to the hypotheses of the experiment, recorded the frequency of actions performed by the infants on the target and test objects. Only the experimenter's back was visible on the videotapes and all sessions were coded with no volume. Thus, the coders could not detect whether the experimenter had demonstrated a target action on an object and could not hear whether objects actually made sounds when actions were performed by either the experimenter or the infants. Thus, we were confident that the coders could not distinguish the surprised, baseline, and predicted conditions from one another.

A detailed coding scheme for each target action was developed for each object set. The target action for the squeaking set was defined by a squeezing motion, that is, the infant gripped and then compressed his/her fingers together on the object (not tapping the object, hitting the object on the table, shaking the object, or gripping it to look at it or passing/throwing it to the experimenter or parent). The target action for the ringing set was defined by a tapping, hitting, or patting motion (not squeezing the object, hitting it on the table, shaking it, or gripping it to look at it or pass/throw it to the experimenter or parent). Finally, the target action for the rattling set was defined by a shaking motion with the wrist and/or whole arm in a back/forth or up/down motion (not tapping the object, squeezing it, hitting the table or a body part with it, or gripping it to look at it or pass/throw it to the experimenter or parent). If the infant performed a fluid shaking movement, then only one target action was counted.

Results

The mean frequency of target actions performed on the different test objects in the surprised and baseline conditions are presented in Table 1[1]. We first examined whether the target properties of the object stimuli were indeed nonobvious to infants by comparing the number of target actions infants performed on test objects after having first seen a functional target object (in the surprised condition) versus a nonfunctional target object (in the baseline condition). We used one-tailed dependent t-tests to compare the frequency of target actions in the surprised condition to those in the baseline condition at each level of shape similarity. (Note that we used one-tailed tests as our predictions were directional). As

[1] In all analyses, we chose not to include the data from the predicted condition as it was difficult to interpret why infants continued to perform target actions on test objects in this condition. That is, it was impossible to distinguish those target actions performed as a result of an expectation about an object's nonobvious property from those performed as a result of the reinforcing nature of the sound property of the test objects themselves (see Baldwin et al., 1993 for a discussion of this issue).

Table 1:Frequency of Target Actions Performed on Test Objects at Each Level of Shape Similarity within each Expectation Condition (Expt. 1).

	Shape Similarity to Target		
Condition	High	Low	Dissimilar
Surprised	2.1 (2.9)	0.6 (1.3)	0.2 (0.9)
Baseline	0.4 (0.7)	0.6 (1.2)	0.0 (0.0)

expected, infants performed significantly more target actions on the high similarity objects in the surprised condition than in the baseline condition, t (19) = 2.39, p < .03. In contrast, infants did not differ significantly in their performance of target actions on the low similarity objects or on the dissimilar objects in the surprised condition versus the baseline condition, (t (19) = .00, p > .99; no statistic was computed for the dissimilar object comparison as no target actions were performed in the baseline condition). These analyses indicated that the appearances of the high similarity objects did not suggest that the objects possessed the nonobvious properties. Instead, infants performed target actions on test objects only after they had been exposed to the properties of particular functional target objects during the testing session.

We next examined the influence of shape similarity on infants' generalization of nonobvious properties within the surprised condition only. As predicted, infants performed significantly more target actions on the high similarity objects than on the low similarity objects (t (19) = 2.55, p < .02), or the dissimilar objects (t (19) = 2.55, p < .02). Furthermore, infants did not differ significantly in their performance of target actions on the low similarity objects and dissimilar objects, t (19) = 1.16, p > .25. The results of these analyses indicate that infants expected objects that shared a high degree of shape similarity to share nonobvious properties, consistent with our hypotheses.

Discussion

As expected, infants performed significantly more target actions on the high similarity test objects in the surprised condition than in the baseline condition. This finding indicates that the appearance of the objects did not suggest the nonobvious properties. Furthermore, in the surprised condition, infants generalized the nonobvious properties to the high similarity test objects but not to the low similarity objects (which still shared textural similarity with the target object), nor to the dissimilar test objects (which differed from the target object in texture, shape, and color). Infants' lack of performance of the target actions on the low similarity objects and on the dissimilar objects indicates that they were not simply imitating any action that the experimenter did--they only imitated the target action when they viewed the test object as a member of the same category as the target object.

The results of this experiment thus indicate that 13-month-old infants will form specific expectations about the nonobvious properties of objects from knowledge gained during the testing session. Furthermore, these findings indicate that infants were relying solely on shared shape similarity to index category membership, an issue discussed further in the General Discussion.

Experiment 2

In Experiment 2, we examined whether 13-month-old infants would treat labels for novel objects as a conceptual marker of object kind and expect those objects that shared the same label to possess the same nonobvious property. The design of Experiment 2 was similar to that of Experiment 1 with one exception: The experimenter labeled the target and test objects with novel count nouns when she introduced them. We predicted that infants would generalize the nonobvious property to objects that shared the same label, even if they shared little shape similarity with the target object.

Method

Participants Participants were 20 infants ranging in age from 12.20 months to 13.85 months (M = 13.02; SD = .54). Ten infants were male and 10 were female.

Stimuli Same as Experiment 1.

Design Same as Experiment 1.

Procedure The procedure was similar to that of Experiment 1, with one exception: The experimenter introduced the target and test objects using novel count nouns (e.g., "Look at this blint!"). Note that the same count noun was used to label the target object and the test objects in a given set (i.e., the target and the test object from the rattling set were all labeled as blints).

Coding Identical to Experiment 1.

Results

The mean frequency of target actions performed on the different test objects in the surprised and baseline conditions are presented in Table 2. As in Experiment 1, we first examined whether the target properties of the object stimuli were nonobvious to infants by comparing the number of target actions infants performed on test objects after having first seen a functional target object (the surprised condition) versus a nonfunctional target object (the baseline condition). (Note we again used one-tailed t-tests as our predictions were directional). As expected, infants performed significantly more target actions on the high similarity objects in the surprised condition than in the baseline condition, t (19) = 2.76, p < .005. Similarly, infants performed more target actions on the low similarity objects in the surprised condition than in the baseline condition, t (19) = 1.86, p < .05. In contrast, infants did not differ significantly in their

Table 2:Frequency of Target Actions Performed on Test Objects at Each Level of Shape Similarity within each Expectation Condition (Expt. 2).

	Shape Similarity to Target		
Condition	High	Low	Dissimilar
Surprised	2.7 (3.9)	2.2 (3.1)	0.5 (1.4)
Baseline	0.3 (0.7)	0.7 (2.3)	0.0 (0.0)

performance of target actions on the dissimilar object in the surprised condition versus the baseline condition (no statistic was computed as no target actions were performed on this object in the baseline condition). These analyses indicated that the appearances of the high and low similarity objects did not suggest that the objects possessed the nonobvious properties.

We next examined the influence of labels on infants' generalization of nonobvious properties within the surprised condition only. In contrast to Experiment 1, infants did not perform more target actions on the high similarity objects than on the low similarity objects (t (19) = 0.51, $p > .60$). However, infants performed significantly more target actions on both the high similarity objects and on the low similarity objects than on dissimilar objects, t (19) = 2.31, $p < .03$ and t (19) = 3.31, $p < .01$, respectively. The results of these analyses indicate that infants expected objects that shared the same label to share nonobvious properties, regardless of shape similarity.

Discussion

In this experiment, infants performed as many target actions on the low similarity objects as on the high similarity objects. This finding indicates that infants relied on shared labels, rather than shape similarity, to guide their inferences about nonobvious object properties, an issue discussed further in the General Discussion. It is important to note, however, that the presence of labels, however, did not lead infants to completely disregard perceptual information. That is, labeling the dissimilar object with the same count noun as the target object did not lead infants to generalize nonobvious object properties to that object. Recall that the dissimilar object shared no perceptual properties in common with the target object. Thus, it appears that some minimal perceptual overlap is necessary for infants to generalize the nonobvious properties.

General Discussion

The present studies were designed to examine the role of object shape similarity and object labels in guiding 13-month-old infants' inferences about nonobvious object properties. The results of our studies yielded three major insights into the nature of infants' inductive reasoning. First, the results of both experiments provide evidence that infants between 12 and 13 months of age will form expectations about shared properties of novel objects after only a ten second experience with a functional target object. Furthermore, infants will extend a specific nonobvious property from a target exemplar to other objects perceived as members of the same category, consistent with the results of previous studies (e.g., Baldwin et al., 1993; Mandler & McDonough, 1996, 1998; Welder & Graham, in press). Thus, our finding that infants could rapidly and efficiently form expectations about the nonobvious properties of novel objects provides important evidence that infants possess well-developed inductive reasoning abilities by the end of the first year of life.

Second, our findings indicate that infants will rely on shape similarity to generalize nonobvious object properties, in the absence of other information about object kind. In Experiment 1, infants were more likely to generalize a nonobvious object property to objects that were highly similar in shape than to objects that were less similar in shape. These findings provide clear evidence that infants expect that objects that share a high degree of shape similarity will also share other "deeper" characteristics. Furthermore, this result suggests that infants appreciate that shared shape similarity is predictive of category membership. That is, infants attend to shape information because it serves as a perceptually-available cue to the underlying structure of a category (see Bloom, 2000; Gelman & Diesendruck, 1999 for a discussion).

Finally, our findings demonstrate that when a novel object is labeled with a novel count noun, infants will overlook shape information and rely on the label to generalize the nonobvious property. In Experiment 2, infants performed as many target actions on low similarity objects as on high similarity objects, when objects were labeled with the same count noun. As discussed earlier, labeling the objects did not lead infants to completely disregard perceptual information as they did not generalize the nonobvious object properties to the dissimilar object (which shared no perceptual features with the target object). Thus, when infants are provided with information about category membership in the form of shared object labels, perceptual information is attenuated. This finding provides clear evidence that young infants can form novel categories and make inductive inferences about nonobvious properties based on a conceptual notion of object kind (see also Mandler & McDonough, 1996, 1998; Welder & Graham, in press). Moreover, our findings indicate that infants as young as 13 months of age recognize the conceptual information conveyed by object labels. That is, infants, like preschoolers, appear to recognize that count noun labels supply information about underlying object kind and furthermore, that members of the same kind share nonobvious properties. This finding is particularly compelling given that the infants in our studies are only just beginning to acquire productive language. Finally, these findings add to a growing body of literature

indicating that naming can foster infants' formation of object categories (e.g., Balaban & Waxman, 1997; Graham, Baker, & Poulin-Dubois, 1998; Waxman, 1999; Waxman & Hall, 1993; Waxman & Markow, 1995) and moreover, these findings provide evidence that labels enhance the inductive potential of categories for young infants.

In summary, the results of these experiments have advanced our understanding of infants' inductive abilities, indicating that 13-month-old infants will use shape similarity and count noun label information for making inferences about nonobvious object properties. These results also suggest a number of important directions for future research. For example, it remains to be seen whether infants who have not yet acquired productive vocabulary (i.e., infants younger than 12 months of age) can rely on object labels to guide their inferences about object properties, and whether count noun labels (versus words from other form classes or versus nonlinguistic stimuli) are privileged in guiding infants' inferences. Research into these issues is currently underway in our lab and we expect that the results of these studies, in conjunction with other recent empirical work, will lead to a coherent account of the developmental processes underlying inductive reasoning during infancy.

Acknowledgments

This research was supported by an operating grant awarded to the first author from NSERC of Canada. The third author was supported by graduate fellowships from NSERC of Canada and the Alberta Heritage Foundation for Medical Research. We thank the parents, children, and adults who participated in the studies as well as the staff at the following Calgary Regional Health Authority clinics for their kind assistance in participant recruitment: Thornhill, North Hill, Ranchlands, and Midnapore. We also thank Kristinn Meagher and Kara Olineck for their assistance with these studies.

References

Balaban, M. T., & Waxman, S. R. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child Psychology, 64,* 3-26.

Baldwin, D. A., Markman, E. M., & Melartin, E. M. (1993). Infants' ability to draw inferences about nonobvious object properties: Evidence from exploratory play. *Child Development, 64,* 711-728.

Bloom, P. (2000). *How children learn the meanings of words.* Cambridge: MIT Press.

Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. *Cognitive Psychology, 20*, 65-95.

Gelman, S. A., & Coley, J. D. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. *Developmental Psychology, 26*, 796-804.

Gelman, S. A., & Diesendruck, G. (1999). What's in a concept? Context, variability, and psychological essentialism. In I.E. Sigel (Ed.), *Development of mental representation theories and applications.* Mahwah, N.J.: Lawrence Erlbaum Associates.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition, 23*, 183-209.

Gelman, S. A., & Markman, E. M. (1987). Young children's inductions from natural kinds: The role of categories and appearances. *Child Development, 58,* 1532-1541.

Graham, S. A., Baker, R. K., & Poulin-Dubois, D. (1998). Infants' expectations about object label reference. *Canadian Journal of Experimental Psychology, 52,* 103-112.

Kalish, C. W., & Gelman, S. A. (1992). On wooden pillows: Multiple classification and children's category-based inductions. *Child Development, 63,* 1536-1557.

Mandler, J. M., & McDonough, L. (1996). Drinking and driving don't mix: Inductive generalization in infancy. *Cognition, 59*, 307-335.

Mandler, J. M., & McDonough, L. (1998). Studies in inductive inference in infancy. *Cognitive Psychology, 37,* 60-96.

Moore, B., & Parker, R. (1989). *Critical thinking: Evaluating claims and arguments in everyday life* (2nd ed.). Mountain View, CA: Mayfield.

Oakes, L. M., Madole, K. L., & Cohen, L. B. (1991). Infants' object examining: Habituation and categorization. *Cognitive Development, 6,* 377-392.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8,* 382-439.

Waxman, S. R. (1999). The dubbing ceremony revisited: Object naming and categorization in infancy and early childhood. In D. L. Medin & S. Atran (Eds.), *Folkbiology*. Cambridge, MA: MIT Press.

Waxman, S. R., & Hall, D. G. (1993). The development of a linkage between count nouns and object categories: Evidence from fifteen- to twenty-one-month-old infants. *Child Development, 64,* 1224-1241.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology, 29,* 257-302.

Welder, A. N., & Graham, S. A. (in press). The influence of shape similarity and shared labels on infants' inductive inferences about nonobvious object properties. *Child Development.*

The Emergence of Semantic Categories from Distributed Featural Representations

Michael J. Greer *(mgreer@csl.psychol.cam.ac.uk)
Maarten van Casteren ^ (maarten.van-casteren@mrc-cbu.cam.ac.uk)
Stuart A. McLellan *(sam26@cam.ac.uk)
Helen E. Moss *(hem10@cam.ac.uk)
Jennifer Rodd * (jrodd@csl.psychol.cam.ac.uk)
Timothy T. Rogers ^ (tim.rogers@mrc-cbu.cam.ac.uk)
Lorraine K. Tyler * (lktyler@csl.psychol.cam.ac.uk)

*Centre for Speech and Language, Department of Experimental Psychology, University of Cambridge
^MRC Cognition and Brain Sciences Unit, Cambridge, UK

Abstract

This paper presents a computational model of semantic memory, trained with behaviourally inspired vectors. The results are consistent with the *conceptual structure account* (Tyler, Moss, Durrant-Peatfield & Levy, 2000), which claims that concepts can be understood, and the effects of random damage predicted, based on (i) the number of correlations its features make, and (ii) the distinctiveness of those correlated features; the former indicating category membership, the latter distinguishing between concepts. The model shows a changing direction of domain-specific deficits as damage accumulates (animals concepts lost first, then objects upon severe lesioning). Also, the pattern of error differs between domains; animals tend to be confused for other members of the same category, whilst object errors disperse more widely across categories and domain. Recent neuropsychological evidence demonstrates a similar pattern for semantically impaired patients. For both patients and the model, this can be attributed to the timing of featural loss: distinctive features are lost earlier than shared features. The model demonstrates that the relative timing of feature loss differs between domains, resulting in the emergence of domain-specific effects.

Introduction

The neuropsychological literature on semantic memory shows patients can develop an impairment in one domain of knowledge, whilst the other is relatively spared. Most commonly, semantically impaired patients show a deficit for living things (e.g. Warrington & Shallice, 1984), with the reverse pattern being rarer (e.g. Hillis and Caramazza, 1991).

There are three main types of explanation for the double dissociation. One postulates physically separate and functionally independent stores in the brain for dissociable categories of knowledge (e.g. Goodglass, Klein, Carey and Jones, 1966; Carramazza & Shelton, 1998). Another suggestion is that concepts may vary by domain according to the type of semantic information upon which they depend, with living things depending more on sensory information and artefacts depending more on functional properties (Warrington & Shallice, 1984; Warrington & McCarthy, 1983; 1987). Selective brain damage to one type of semantic information will lead to a category-specific deficit. This account assumes neuro-anatomical specialisation for type of property rather than category per se, to permit their independent disruption by brain damage. Finally, and most recently, attempts to account for category-specific deficits suggest that they can emerge from the *internal structure of concepts* alone without any type of neural or functional specialisation. Computational models have shown that random damage to a unitary, distributed system can produce category-specific deficits (e.g. Devlin, Gonnerman, Anderson and Seidenberg, 1998; Tyler et al, 2000). These models draw on structural aspects such as property correlation and distinctiveness.

Conceptual Structure Account

Common to all distributed accounts of semantic memory (see McRae, de Sa, Seidenberg, 1997; Devlin et al, 1998; Tyler et al, 2000) is the observation that similar concepts tend to have overlapping sets of semantic features. Properties that frequently co-occur in concepts will serve to predict each others presence, a fact that a distributed connectionist network will exploit during training, leading to mutual activation of those properties. A consequence of mutual activation is resilience to damage of those properties, and hence their continued 'availability' to a stricken network when identifying concepts. A second important factor, is the distinctiveness[1] of features (cf. Devlin et al, 1998). A feature that is present in only one concept can be used

[1] Distinctiveness is calculated as 1/number of concepts for which the property is given.

to discriminate that concept from all others. As a feature occurs in an increasing number of concepts it becomes a progressively poorer marker for each of those concepts.

The *conceptual structure account* of semantic memory (Durrant-Peatfield, Tyler, Moss, & Levy, 1997; Tyler et al 2000) recognises that these factors – correlation and distinctiveness – will interact to determine which features will survive random damage, and the usefulness of the remaining features in preventing concept loss. By their very nature correlation and distinctiveness tend to be inversely related with each other. Highly correlated properties are often present in many concepts, and hence are not very distinctive. Thus, they will be robust to damage, but their preservation will be more useful for identifying the category to which an item belongs rather than distinguishing it from other category members. However, those distinctive properties that do correlate with other properties (especially other distinctive properties) will protect the concept in which they are found. Distinctive properties that fail to make strong correlations with other properties will be very vulnerable to damage. Domain differences and dissociations arise because concepts in different domains differ in these respects. We theorize that living things concepts have many intercorrelated properties, compared to artefacts, but these tend to be less distinctive. As a consequence, artefact concepts are more robust at all but severe levels of damage when only highly correlated properties remain intact.

Computational Model

Previous work instantiating the conceptual structure account of semantic memory (Tyler et al, 2000) used 16 vectors that incorporated the theoretical assumptions of the account. In the current model the vectors are designed to broadly reflect the observed differences between living and non-living domains, as found in a large-scale property generation study (Moss, Tyler & Devlin, In Press). The simplified vectors, homogenous within domain and of equal number between domains, ensure the model's results are readily interpretable. In addition, the model was scaled up, and trained on 96 vectors. Consequently, the training set is as sparse as the property norm data which it resembles. One might expect a distributed model to perform differently as the training set becomes more sparse, as a sensible error-reduction strategy would be to turn all units off. This model sought to confirm that a distributed model would still build internal representations reflecting correlational structure in spite of extreme sparsity.

Property norm data

Tables 1 and 2 report the global and distributional statistics of the property norm concepts. Data is also given for the model vectors, designed to resemble the property norm concepts as far as practicable.

Table 1: Global properties of the property norm set and the model vectors

	Property norms	**Model vectors**
Number of concepts	93	96
Features that are highly distinctive[2]	78%	78%
Sparsity[3]	3.7%	4.6%

Table 2: Characteristics of property norm concepts across domains (figures for model vectors in brackets)

	Living things	**Artefacts**
Mean no. properties/concept	17.7 (20)	11.3 (14)
Mean distinctiveness of properties	0.64 (0.22)	0.73 (0.32)
No. of shared properties/concept[4]	13.7 (15)	7.5 (6)

Following McRae et al (1997) the Pearson product moment correlation was computed for all pairs of semantic features. For the property norms, of the 78,210 possible correlations, only 2332 scored $|r|>0.3$. Living things had more correlated property pairs (CPPs) than artefacts (4070 vs. 1612), but artefacts had proportionally more CPPs occurring between distinctive features (20.0% artefacts vs. 11.7% living things).

Representations

The training set consisted of 96 vectors, divided into 2 domains (Animals and Objects) and 4 categories (labelled somewhat arbitrarily as Land animals, Birds, Tools and Furniture). The vectors embodied the facts outlined below:

- There were 48 Animal vectors (24 Land animals and 24 Birds).
- Each Animal vector turned on 20/368 features.
- Every Animal vector turned on 10 'Animal shared' features.

[2] The proportion of features in the set shared by just 1 or 2 concepts.

[3] Sparsity refers to the average proportion of features turned on for each vector.

[4] A shared property being defined as one held by three or more concepts, otherwise the property is distinctive.

- Land animals were distinguishable from Birds by which group of 5 shared features was turned on – 'Land shared' or 'Bird shared'.
- All animals could be distinguished from each other by which three 'Animal distinctive features' were on.
- There were 48 cross-domain features, each concept turning on two. Each concept, whether animal or object turned on a unique combination of cross-domain features (say 1 and 4; or 2 and 5 etc) such that each unit was turned on by 4 different concepts (2 animal, 2 objects). This means that having a cross-domain feature on does not predict at all which domain's concept is on, but limits the concept to one of four possibilities.
- There were 48 object concepts (24 tools and 24 furniture).
- Every Object vector turned on 14/368 features.
- Tools and Furniture were distinguishable by which group of 6 shared features they turned on.
- Object concepts were identifiable by which 2 'object distinctive triplets' were on. They are termed triplets because, within a triplet, if one feature is on then the other two must also be on (likewise when off). However, each "object distinctive triplet" is turned on both by 1 tool concept and 1 furniture concept, so having a triplet on does not perfectly predict which object is on (in contrast to "animal distinctive features").

The resultant vectors resemble the concepts analysed in the property generation study (see tables 1 and 2). The resemblance extends to the vectors' correlational structure. For the model vectors, of the 67,528 possible correlations, 1864 scored $|r|>0.3$. Animals had more correlated property pairs (CPPs) than objects (5472 vs. 2016), but objects had proportionally more CPPs occurring between distinctive features (35.7% objects vs. 2.6% animals). This reproduces the pattern of the domain effects in the property norms, but exaggerates the size of the difference.

Architecture and training

The network consisted of three layers, a semantic input layer, a hidden layer and a semantic output layer, as shown in Figure 1. During training, with the back-propagation learning algorithm (Rumelhart, Hinton & McClelland, 1986), the network was required to reproduce the input on the output layer. 10 networks were trained with different initial random weights (±0.005), with a learning rate of 0.25 and momentum 0.5. Training was stopped when the squared error for each feature in every vector was below 0.01, occurring after a mean of 193 presentations of the complete target set.

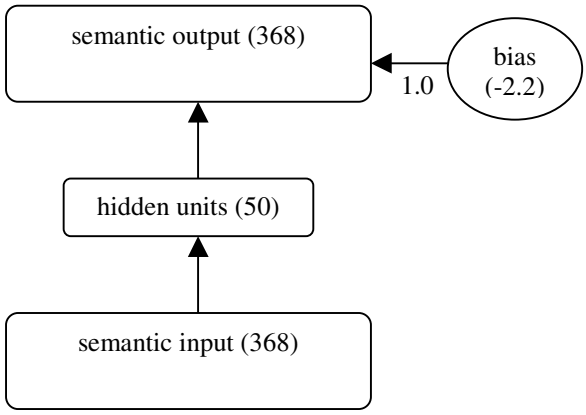


Figure 1: Model architecture: the numbers in each box indicate the number of units in that layer, while arrows indicate full connectivity between layers[5]

Lesioning

Brain damage is simulated in this model by random deletion of semantic connections (by setting weights to 0). Initially 10% of weights were cut, then the model's performance analyzed. The proportion of damaged connections was increased by increments of 10% until all inter-layer connections were set to 0. This lesioning process was carried out 5 times on each of the 10 trained networks to produce a total of 50 networks.

Testing

Network performance was analyzed both at the level of features and concepts. The training set was presented to the network's input layer and the pattern of activations on the output layer examined. We predict that highly shared features will be more resilient to damage than distinctive features (i.e. will still activate when they should). In the model vectors this will correspond to the greatest advantage being for 'animal shared features', then 'land', 'bird', 'tool' and 'furniture' shared features behaving similarly, with 'animal distinctive' features being least preserved. A different pattern is predicted at the conceptual level where discrimination is dependent primarily upon distinctive features. Object distinctive features cluster into 'triplets', this additional inter-correlation is predicted to enhance their robustness to damage relative to Animal distinctive features, leading to an advantage in concept identification. Only at severe levels of damage, when all distinctive features are lost to the network, will the advantage for animal shared features translate to an advantage in concept naming.

[5] Bias ensures the 100% damaged model outputs 0s (approximately); in its absence the semi-linear logistic activation function makes every semantic output unit 0.5. Bias connections were not lesioned.

Featural analysis

For each vector, the activation of the semantic output layer was binarised – unit values <0.5 were scored 0 while values ≥0.5 were scored 1. Each unit value was compared to that of the input vector and declared correct or error. Attention focused on the subset of units that were supposed to be on for each vector, and the number of errors summed for each domain. Because each output unit represents a local feature it is possible to compare the errors across the different feature types (i.e. 'animal shared', 'object distinctive triplet' etc).

Overall analysis

The pattern of activity over the output units was compared to all 96 vectors; both had been normalised to remove effects of concept size[6]. The closest match, by Euclidean distance, was considered the model's response. Upon network lesioning, errors occur of three types: within-category error, cross-category error, or a cross-domain error.

Results

Figure 2 presents the results of the featural analysis averaged over 50 simulations. The general pattern is for a greater impairment of unique relative to shared features at all levels of lesioning between 20 and 80%.

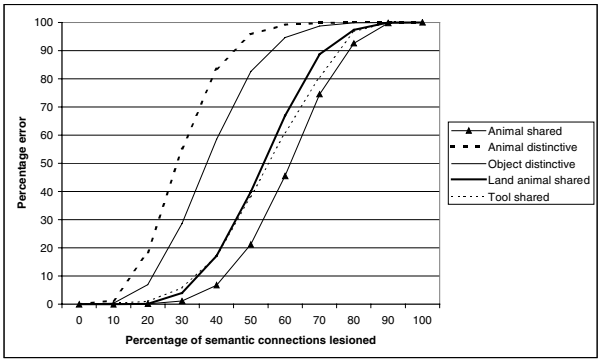


Figure 2:Featural error (failure to activate) as a function of network damage.

The effects of damage upon concept identification are shown in Figure 3, which shows an advantage for objects over animals up until 80% of connections had been lesioned. A two-way ANOVA (domain*damage) showed a main effect of domain (i.e. animals vs. objects, (F[1,539]=317, p<.0001), and a significant domain by damage interaction (F[10,539]= 5406, p<.0001).

A repeated measures t-test on the 90% damaged data-points showed that the advantage for animals, although small, was significant (t= -6.249, p<.0001).

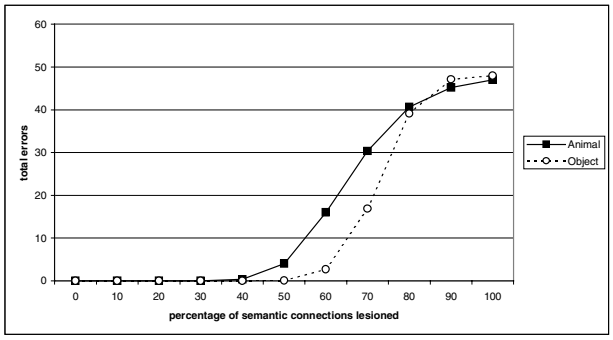


Figure 3: Identity mapping as a function of damage.

Figure 4 shows the difference in the distribution of error types when concepts were mis-identified. It attributes the early animal deficit to within-category error, with all errors involving members of the same category. Conversely, object errors are more widely dispersed between the two domains.

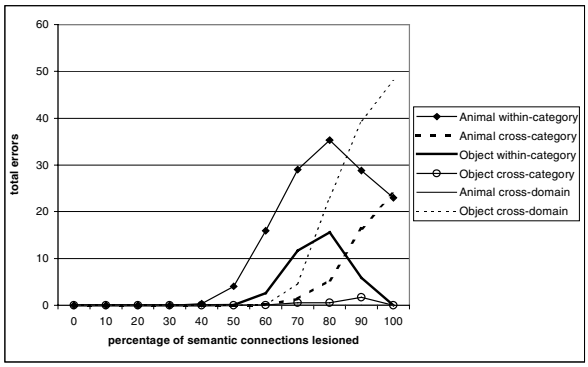


Figure 4: Error types as a function of damage.

Discussion

This research demonstrates how category and domain-specific deficits can arise following damage to a single distributed semantic system without explicit category structure. Further, it accounts for the patterns of impairment observed in patients as resulting from a complex interaction of correlations between features and the extent to which features are shared or distinctive. In general, the behaviour of the model is very similar to the behaviour of some brain-damaged patients with category/domain specific deficits.

The preservation of individual features was dependent on the number of correlations each feature entered into, with more highly shared, correlated features being more robust. This pattern is similar to that observed in semantically-impaired patients who show better preserved knowledge of shared, category-

[6] Concept size refers to the number of features turned on.

defining information compared to distinctive, concept-identifying information (Moss, Tyler, Durrant-Peatfield, & Bunn, 1998; Moss et al, In Press).

The preservation of individual concepts showed a different pattern. Global damage, where connections between layers were randomly lesioned, produced an initial impairment for animals, followed at severe lesioning, by impairment for objects. Successfully identifying a concept relies most heavily on activating its distinctive features. With damage, object distinctive properties were more robust than animals, which can be attributed to their tendency to correlate with other distinctive properties. Crucially, this same pattern of correlations occurred in the empirically derived property norms (Moss et al, In Press). This shows that the same factor that accounts for an early animal deficit in the computational model could also account for the initial living-things deficit found in at least some semantic dementia patients (Moss & Tyler, 2000; Moss et al, In Press). Beyond 70% lesioning all distinctive features failed to activate, whether animal or object. Consequently, the model had to 'make a guess', though the odds of guessing correctly would have differed for the two domains. The animal shared features are both more numerous and more correlated, hence will remain more likely to be available to the network. Therefore the model would have been guessing from a smaller subset of possible concepts than would have been the case for objects, which could lead to a mild object deficit. This unequal distribution of shared features between domains is also characteristic of the property norm data.

The lesioning data shows that the magnitude of the early disadvantage for naming animals exceeds that of the late disadvantage for naming objects. This too seems to be reflected in semantic dementia patients where living-thing deficits tend to be both greater in size and more numerous than corresponding artefact deficits (Moss & Tyler, 2000).

The conceptual structure account, realised in the computational model, also predicts the type of error likely to be made when concepts are mis-identified. Due both to the robustness of animal shared features, and the vulnerability to damage of their distinctive features, animals will most commonly be confused with other members of the same category. Cross-domain errors should hardly ever occur. Whilst the same should be true of objects, this tendency will be less marked, and errors will be dispersed more widely between the types of error possible: within-category; cross-category; and cross-domain. There is some evidence for this pattern in longitudinal studies of picture naming. Hodges, Graham & Patterson (1995) report a semantic dementia patient, JL. For living things, he made progressively more within-category and superordinate errors, but never produced a cross-category or cross-domain error. Similarly for the progressive aphasic patient AA (Moss et al, 1999), tested on four occasions over two years, but failing to produce a living things cross-category mistake until the final testing session. For artefacts, in contrast, she occasionally made cross-category and cross-domain errors throughout the testing period.

Epilogue: The problem of determining error

In common with other models of semantic memory, the network's output was compared to every vector in the training set, and the vector with which the normalised Euclidean distance was smallest, regarded as the network response. As a result the network was forced to make a response, irrespective of the meaningfulness or otherwise of the output. A potential limitation of this procedure is that the network could not respond "don't know", a response commonly produced by semantically impaired patients in semantic tasks. As a first step to simulating a "don't know" response a threshold for normalised Euclidean distance was introduced; if the error between the output and every vector exceeded this threshold then the output was scored incorrect. The problem then was to decide how strict the threshold ought to be. The result of some early explorations is shown in figure 5.

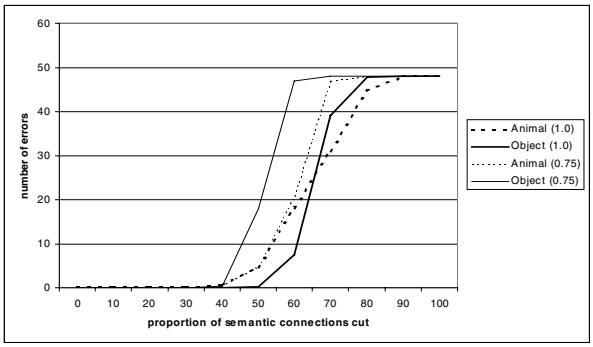


Figure 5: The variability of domain effects with different thresholds for normalised Euclidean distance.

A threshold value of 1.0 produced a strong cross-over, an early animal deficit progressing to an object deficit with damage. Reducing the threshold value to 0.75 had a catastrophic effect upon object identification, but animal identification was less impaired. As a result, the graph showed a consistent deficit in identifying objects. A threshold figure of 1.5 produced a graph identical to that when no threshold was applied (i.e. the same as figure 3). The sensitivity of object identification to varying threshold values probably reflects the smaller number of inter-correlations between object features. The representation of an object concept in semantic space will be sparser, so when the representation is damaged, and a strict threshold is applied, it will be unlikely to fall into a neighbouring concept's space.

Instead the output will be scored "don't know", an outcome which is much less likely in the denser animal semantic space.

The problem here is how we determine where the error threshold should lie? One approach might be to record the number of "don't know" responses the model makes, and relate this to patient data. This is complicated by the difficulty of relating the degree of brain damage to particular levels of network lesioning. Also, patient performance on tests of semantic knowledge varies with the demands of the task. For example, "Don't know" is a more common response for picture naming than word-picture matching.

Speculatively, the normalised Euclidean threshold could reflect a compromise position in a speed/accuracy trade off. Tasks that demand a rapid response would have a more relaxed threshold than those where time is given for a considered response.

Conclusion

This model suggests that the data inherent in conceptual structure is sufficient to account for the domain-specific effects observed in semantically impaired patients. Categories and domains emerge when concepts are represented in a single, distributed system. Some recent neuro-imaging studies fail to show regional differences in activation for different conceptual domains (e.g. Devlin, Russell, Davis, Price, Moss, Matthews, & Tyler, 2000), consistent with the neural substrate of concepts being organised in a distributed fashion.

Acknowledgements

Grants from the McDonnell-Pew Foundation, Medical Research Council (UK) and the Wellcome Trust supported this research.

References

Carramazza, A.C., & Shelton, J. R. (1998) Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of Cognitive Neuroscience*, 10, 1-35.

Durrant-Peatfield, M., Tyler, L.K., Moss, H. E. & Levy, J. (1997) The distinctiveness of form and function in category structure: A connectionist model. In: M.G. Shafto & P. Langley (Eds.), *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, Stanford University, Mahwah, NJ: Erlbaum.

Devlin, J., Gonnerman, L., Anderson, E., and Seidenberg, M. (1998). Category specific semantic deficits in focal and widespread brain damage: A computational account. *Journal of Cognitive Neuroscience*, 10, 77-94.

Devlin, J.T., Russell, R.P., Davis, M.H., Price, C.J., Moss, H.E., Matthews, P., & Tyler, L.K. (2000) Susceptibility-induced loss of signal: Comparing PET and fMRI on a semantic task. *NeuroImage*, 11, 589-600, 2000

Goodglass, H., Klein, B., Carey, P., & Jones, K. (1966). Specific semantic word categories in aphasia. *Cortex*, 2, 74-89.

Hillis, A. E., & Caramazza, A. C. (1991). Category-specific naming and comprehension impairment: A double dissociation. *Brain & Language*, 114, 2081-2094.

Hodges, J., Graham, N. & Patterson, K. (1995) Charting the progression in semantic dementia: Implications for the organisation of semantic memory. *Memory*, 3, 463-495.

McRae, K., de Sa, V., & Seidenberg, M. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, 126, 99-130.

Moss, H. E., & Tyler, L. K. (2000) A progressive category-specific deficit for non-living things. *Neuropsycholgia, 38, 60-82.*

Moss, H. E., & Tyler, L. K. (1997) A category-specific semantic deficit for non-living things in a case of progressive aphasia. *Brain and Language*, 60, 55-58.

Moss, H. E., Tyler, L. K., & Devlin, J. (In Press) The emergence of category-specific deficits in a distributed semantic system. In E. M. E. Forde and G. W. Humphreys (Eds.) *Category-specificity in mind and brain*.

Moss, H. E., Tyler, L. K., Durrant-Peatfield, M., & Bunn, E. (1998) "Two eyes of a see-through"; Impaired and intact semantic knowledge in a case of a selective deficit for living things. Neurocase, 4, 291-310.

Rumelhart, D. E., Hinton, G. E. & McClelland, J. L. (1986) A general framework for parallel distributed processing. In D. E. Rumelhart and J. L. McClelland (Eds.) *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: MIT Press.

Tyler, L.K., Moss, H. E., Durrant-Peatfield, M., & Levy, J. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain & Language, 75, 195-231.*

Warrington, E. K., & McCarthy, R. (1983). Category specific access dysphasia. *Brain*, 106, 859-78.

Warrington, E. K., & McCarthy, R. (1987). Categories of knowledge: further fractionations and an attempted integration. *Brain*, 110, 1273-96.

Warrington, E. K., & Shallice, T. (1984). Category specific impairments. *Brain*, 107, 829-54.

Beliefs Versus Knowledge: A Necessary Distinction for Explaining, Predicting, and Assessing Conceptual Change

Thomas D. Griffin (tgriffin@uic.edu)
Stellan Ohlsson (stellan@uic.edu)
Department of Psychology, 1007 West Harrison Street (M/C 285)
Chicago, IL 60607, U.S.A.

Abstract

Empirical research and theoretical treatments of conceptual change have paid little attention to the distinction between knowledge and belief. The distinction implies that conceptual change involves both knowledge acquisition and belief revision, and highlights the need to consider the reasons that beliefs are held. We argue that the effects of prior beliefs on conceptual learning depends upon whether a given belief is held for its coherence with a network of supporting knowledge, or held for the affective goals that it serves. We also contend that the nature of prior beliefs will determine the relationship between the knowledge acquisition and the belief revision stages of the conceptual change process. Preliminary data suggests that prior beliefs vary in whether they are held for knowledge or affect-based reasons, and that this variability may predict whether a change in knowledge will result in belief revision.

Introduction

Theorists and researchers tend to agree that prior concepts often impede people's ability to learn conflicting information (e.g., Chi, 1992; Dole & Sinatra, 1998; Thagard, 1992). The paradox of knowledge acquisition is that new information can only be understood in terms of existing ideas, yet existing ideas act as a filter, often distorting new information to make it more consistent with prior concepts. This raises the question of how we ever learn anything fundamentally new.

At least two widely cited models agree about crucial steps in the conceptual change process (Chi, 1992; Thagard, 1992): (1) recognizing that the new information conflicts with (or is fundamentally different from) existing concepts; (2) constructing a new knowledge structure to support the new information; (3) replacing the old concepts with the new, more coherent concepts. In short, conceptual change involves learning new concepts and then substituting them for the old. Though these models (i.e., Chi, 1992; Thagard, 1992) diverge in some of the details about the process, they seem to agree that conceptual coherence largely determines conceptual replacement.

In addition, Ohlsson and Lehtinen (1997) have suggested that the use of abstract schemas may be needed to explain how new knowledge representations can be created that will not be distorted by the conflicting prior concepts. They suggest that activation of conflicting prior concepts activates related abstract concepts that are not in direct conflict with the new information, and that can be utilized in constructing an accurate representation of the new information. These aspects of the conceptual change process become more important when we consider how beliefs differ from knowledge.

Belief Versus Knowledge

The present paper defines knowledge as the comprehension or awareness of an idea or proposition ("I *understand* the claim that humans evolved from early primates"). After a proposition is known, one can accept it as true ("I *believe* the claim that..."), reject it as false ("I *disbelieve* the claim that..."), or withhold judgment about its truth-value ("I have *no opinion* about the claim that...").

The present knowledge/belief distinction is intended to be psychological. Thus, knowledge and belief refer to qualitatively different aspects of the mental representation: knowledge refers to the representation of a proposition, and belief refers to the representation of a truth-value associated with a proposition. These definitions are consistent with Quine, and Ullian's (1970) argument that people can have knowledge of an idea or proposition, but either not believe it to be true, or hold a belief that the concept is false.

A proper distinction between comprehension versus acceptance or rejection of an idea allows us to consider the multiple influences on belief formation, and to speak more clearly (and realistically) about the relationship between knowledge and belief change.

A Change in Knowledge, Beliefs, or Both?

It is striking that discussions of conceptual change use the terms 'knowledge', 'beliefs', and 'prior conceptions' interchangeably. Recently, diSessa (2000) highlighted the unexamined lack of agreement among conceptual change researchers regarding basic issues such as whether the conceptual change refers to a change in concepts, beliefs, nodes, or links.

In the models previously mentioned (Chi, 1992; Thagard, 1992) conceptual change involves both the creation of new knowledge (step 2), and a process of abandoning the old ideas in favor of the new ones (step 3). Step 2 amounts to acquiring new knowledge or conceptual understanding, while step 3 amounts to belief revision. Thus, belief revision is characterized as rationally disavowing a prior belief whenever the computed conceptual coherence of a new knowledge structure is higher. This account presupposes that knowledge is the only foundation for belief.

The present argument favors a distinction between knowledge and belief, but acknowledges that knowledge of a concept must precede any judgment of its truth or falsehood. The issue being raised here is that acceptance or rejection of a concept may not be solely contingent upon the coherence of its relations to supporting knowledge in the form of evidence, argument, and logical implications. Belief in a concept may serve affective and social functions. Thus, people might accept a certain idea independent of its coherence with relevant knowledge, and perhaps change a belief even though it will reduce conceptual coherence. Beliefs may vary qualitatively in the degree to which they are part of a specified and coherent network of relevant knowledge.

This claim about the varying bases of belief is similar to one made regarding the affective versus cognitive bases of attitudes (e.g., Eagly & Chaiken, 1993). However, while attitudes refer to subjective evaluations of objects as 'positive' or 'negative', beliefs refer to the acceptance or rejection of propositions. Knowing that a person believes in a proposition, such as "humans evolved form primates", tells us nothing about whether that person feels positively or negatively about this state of affairs. Furthermore, attitude-change theories predict greater change in attitudes that are grounded more in affect than cognitions: precisely the opposite prediction being made here for affect-based beliefs.

A number of implications arise when considering the variability in the bases of beliefs. This variability could influence how belief-conflicting information is processed and understood. Even if a new conceptual scheme is understood, the role of affective motivations in belief acceptance calls into question the notion that belief revision results every time a new conceptual scheme increases coherence.

Belief Bases and Conceptual Learning

Part of the importance in making a knowledge/belief distinction lies in the potential influence that a belief's underlying knowledge structure (or lack thereof) might have on the comprehension of belief-conflicting information. The greater conceptual coherence of knowledge-based beliefs should aid in conceptual learning by making conflict recognition more likely and providing a context that facilitates the construction of a new conceptual representation.

Knowledge-based beliefs are defined by greater coherence with related conceptual networks than beliefs held for affective goals. Therefore, a person should more easily recognize when the evidence, argument, or logical implications of a new conceptual scheme are in conflict with a knowledge-based belief than with an affect-based belief. As already stated, conflict recognition is the first step in the conceptual change process and a necessary step in avoiding distortion via assimilation (Chi, 1992; Thagard, 1992). Thus, when people encounter information that conflicts with knowledge-based versus affect-based beliefs, they should be less likely to assimilate and distort and more likely to begin the process of constructing an accurate representation of the information.

Following conflict recognition, the richer and more detailed conceptual context of knowledge-based beliefs should help compare and contrast the belief with the new information, thus facilitating the process of constructing a representation of the new information. By definition, a prior belief that conflicts with new information is relevant to that new information. The concepts of ‘black’ and ‘white’ contain components that oppose one another, yet our understanding of ‘black’ seems to rely heavily on its contrast with the concept of ‘white’. Conceptual contrast could help highlight the boundaries that separate and therefore define the concepts.

In addition, contrasting old and new concepts might lead to the generation of useful abstractions that capture the principles underlying the contrast. In fact, the conceptual network underlying knowledge-based beliefs may provide direct links to abstract concepts that already exist. Given the theorized role of abstract concepts in constructing new representations (i.e., Ohlsson & Lehtinen, 1997), the conceptual framework provided by knowledge-based beliefs could prove quite beneficial in the comprehension of conflicting concepts.

The hypothesis that conflicting beliefs can assist learning via conceptual contrast has received indirect support from classroom studies of pedagogical techniques (for a review, see Guzzetti, Snyder, Glass, Gamas, 1993). A meta-analysis of 70 reading and science education studies revealed that techniques that contrasted new concepts with common misconceptions resulted in better comprehension of the new concepts compared to a number of alternative techniques. The problematic conclusion drawn from was that prior concepts impede learning, so they must be refuted. However, refutational techniques did not directly refute the students' *own* prior concepts, but rather informed students of *common* misconceptions, then presented new concepts as a contrast to these misconceptions.

It is possible that merely highlighting contrasting concepts facilitated comprehension. This interpretation is consistent with the fact that students who did not already possess the prior concepts appeared to benefited equally from refutation compared to students who did have the misconceptions. Contrary to a common assumption, these results could be evidence that conflicting prior beliefs can aid in conceptual learning, so long as the prior concepts are made salient and explicitly contrasted with the new information.

In sum, there are sound theoretical reasons to expect that knowledge-based beliefs should lead to greater conceptual understanding of conflicting information than affect-based beliefs. In fact, if the initial problem of recognizing conceptual conflict is overcome, then the conceptual framework of knowledge-based beliefs may result in greater comprehension than when there are no prior beliefs at all. However, the final stage of conceptual change (i.e., belief revision) remains.

Belief Bases and Belief Revision

There are some obvious reasons to expect that the third and final stage of conceptual change will also be influenced by the underlying source of prior beliefs.

Updating beliefs with new knowledge should be heavily influenced by motivation and epistemological values. Those who are affectively motivated to form beliefs independent of conceptual coherence will have little motivation to revise those beliefs in light of new ideas that could increase coherence. These people may be specifically motivated to 'isolate' new information and actively avoid evaluative comparisons of conceptual coherence. Recent work has shown that different kinds of beliefs are associated with different epistemological values, and these values predict how different beliefs are affected by anomalous information (Chinn & Brewer, 2000). Conceptual replacement or belief revision may follow the competitive rules of conceptual coherence, but only when the initial belief is based upon its coherence with other knowledge.

Beyond motivational influences on belief revision, the conceptual structure of prior beliefs is likely to have direct cognitive effects on coherence comparisons. Making meaningful comparisons between affect-based beliefs and new conceptual knowledge will prove difficult given the different levels of conceptual specification. Also, if coherence comparisons are made, the lack of conceptual specificity inherent in affect-based beliefs will make any revision short of complete abandonment cognitively difficult. Thus, affective beliefs seem to face an 'all or nothing' dilemma, where the most probable outcome is a lack of belief revision. Issues of motivation and conceptual structure make it unlikely that affect-based beliefs will be revised following the comprehension of a coherent conceptual framework.

In short, the present paper advances the following arguments: (1) beliefs differ from knowledge; (2) beliefs vary in whether they are held for coherence with supporting knowledge versus affective motives; (3) conceptual change involves both knowledge acquisition and belief revision; (4) the variability in prior belief bases may influence both of these components of the conceptual change process. The first step in the empirical validation of these claims is to demonstrate that people believe in different concepts for affective as well as knowledge-based reasons and that this difference is related to their willingness to change those beliefs.

Method

The following study was a preliminary investigation of the variability in the underlying bases for beliefs. Any attempt to assess whether beliefs are the result of knowledge coherence or affective goals will have weaknesses that can only be overcome with the use of multiple converging methodologies. Our modest goal in this study was to examine whether people would self-report that their beliefs were held for largely affective reasons and not due to support from relevant knowledge. We hoped to demonstrate variability across beliefs regarding their bases in knowledge versus affect. We also expected that people's underlying reasons for holding their beliefs would be related to their self-reported willingness to revise those beliefs in the face of strong conflicting evidence.

Participants

Participants were 120 undergraduates at the University of Illinois at Chicago.

Belief Assessment Materials and Procedures

Participants reported their prior beliefs on five different topics: creationism, evolution, extra sensory perception, the 'opposites attract' theory of romantic attraction, and the existence of an afterlife. The topics were chosen for their potential relationships to relevant knowledge and affect. For each topic participants were told "for the purpose of this study [topic] is defined as...", followed by a one sentence description of the topic. The descriptions were worded as simply as possible and participants were told to ask for clarification if needed.[1]

Following each description was the question "To what extent do you believe in [topic]?" Participants indicated their level of belief on a scale ranging from 1 (completely disbelieve) to 9 (completely believe).

[1] No participants asked for clarification. Also, failures in definition comprehension would be consistent with the present theory, and would not alter the interpretation of the present results.

Participants were then asked to list their top 3 reasons for their belief on the topic. This self-generation task was designed to examine participants' most accessible reasons that they had previously associated with their belief on the topic.

After participants reported their degree of belief and listed their reasons for all five topics, they were presented with a list of potential reasons why a person might hold any given belief. They were told "for each potential reason below indicate whether that reason is why you personally, hold your belief about the idea." Participants responded on a likert-scale ranging from 1(not at all my reason) to 9(completely my reason). Five reasons were presented, two knowledge based reasons and three affective reasons (see Table 1).

Table 1. Knowledge and Affective Reasons for Belief.

Affective Reasons*:
My belief about [topic] makes me feel good or is comforting.
When it comes to issues like [topic], I trust my 'heart', not my 'head' to tell me the truth.
I don't need proof, I have faith that my belief about [topic] is correct.
Knowledge Reasons*:
My belief about [topic] is a result of examining all of the evidence I'm aware of and choosing the most convincing explanation.
My belief about [topic] is supported by current scientific knowledge.

*Reasons were not labeled 'affective' or 'knowledge'.

After rating their reasons for each belief, participants were asked: "Imagine that you were presented with strong evidence that contradicted your belief. How likely would you be to change your belief?" Participants indicated their willingness to change each belief on a scale from 1(not at all) to 9(completely).

Results

For each of the five topics, participants were classified into one of three groups: 1 to 3 rating = 'disbeliever'; 4 to 6 rating = 'no opinion'; 7 to 9 rating = 'believer'. Believers on one topic could be believers, disbelievers, or have no opinion on the other topics. The results that follow compare the knowledge versus affective reasons given by believers and disbelievers across the five topics. Those with no opinion are not included here.

Self-Generated Reasons

A qualitative examination of participants' self-generated reasons for belief revealed that participants gave both affective and knowledge-based reasons for their beliefs. The types of reasons varied systematically across topics and between believers and non-believers (see Table 2, for prototypical reasons given by 'believers'). The reasons given for belief in creationism, an afterlife, and disbelief in evolution were rarely knowledge-based and often referred to affect. Some participants simply mentioned that the belief was part of their religion or just how they were brought up. In contrast, belief in evolution, and disbelief in creationism and an afterlife were never supported with affective reasons and participants often referred to evidence. Belief and disbelief in ESP and opposites attract was most often supported by personal experience or reference to media portrayals, but affective reasons were also provided.

Table 2. Prototypical reasons for belief*.

Creationism
"I rely on faith"; "the bible says so"; "I couldn't live if I didn't think there was a God"
Afterlife
"I hope there is one"; "It relieves my fear"; "life would be meaningless otherwise"
Evolution
"biological evidence"; "You can observe similarities between species"
ESP
"I have this ability"; "t.v. documentaries"; "It sounds cool"
Opposites attract
"personal experience"; "media"; "I have seen it"

*Reasons for disbelief are not included in this table

Ratings of Knowledge Versus Affective Reasons

The mean scores were calculated for the three affect reasons and the two knowledge reasons. Figures 1 and 2 report the mean levels of knowledge and affect scores, and self-reported willingness to change a belief or disbelief in the face of conflicting evidence. As mentioned previously, each participants belief or disbelief varied as a function of topic. This presents a problem for any straightforward statistical test of the Belief X Topic interaction. Thus, a qualitative comparison of the means was followed-up by a formal test using correlational methods.

As expected, both the mean levels of affect and knowledge associated with a belief varied across the different topics (see Figure 1). Belief in evolution was associated with higher knowledge than affect scores, and with lower affect scores than belief in the other four topics. In contrast, belief in creationism and an afterlife were associated with higher affect than knowledge scores, and with higher affect than belief in the other three topics.

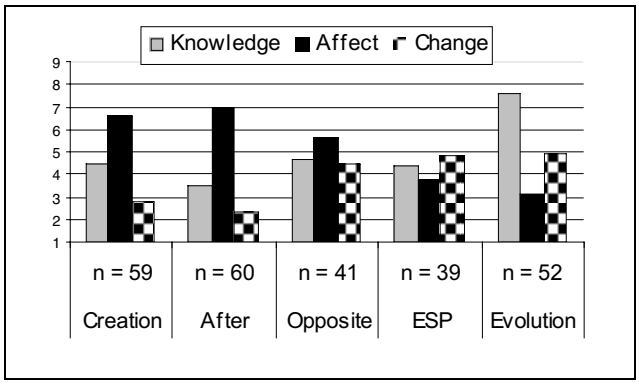


Figure 1: Reasons and Will to Change for Believers.

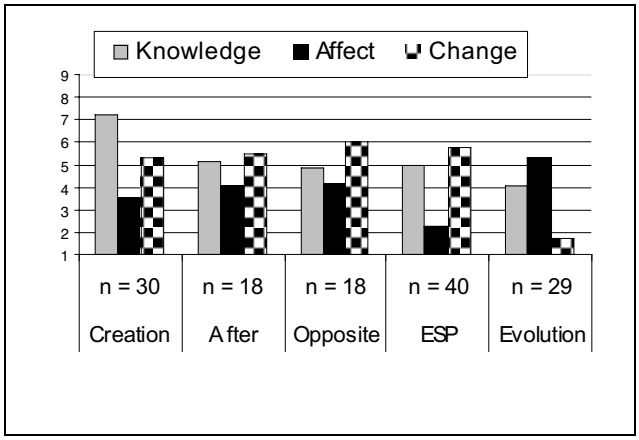


Figure 2: Reasons and Will to Change for Disbelievers.

The opposite pattern of results was found for disbelief across topics (see Figure 2). Specifically, disbelief in creationism was associated with higher knowledge than affect scores, and higher knowledge scores than disbelief in the other four topics. In contrast, disbelief in evolution was associated with higher affect than knowledge scores, and higher affect than the other topics. Overall, the differences in knowledge and affective scores for belief and disbelief across the five topics were consistent with the kinds of reasons participants provided in the self-generation task.

To get around the problem of the membership overlap between believers and disbelievers across topics, belief was treated as a continuous variable. Also, a single score of endorsed reasons was calculated for participants on each topic, by subtracting their affect score from their knowledge score on each topic. Each of the five difference scores was correlated with the continuous measure of belief for the corresponding topic. The results are the five coefficients in Table 3.

For all topics, the knowledge-affect difference score was significantly correlated with the tendency to believe or disbelieve. Negative correlations suggest that as belief in the issue was greater, knowledge reasons decreased compared to affect reasons. This was the case for belief on all issues except for evolution. As belief in evolution increased the endorsement of knowledge reasons increased greatly compared to affect reasons. These findings are consistent with the descriptive examination of the mean scores for believers and disbelievers.

Reasons for belief were also related to self-reported willingness to change. As seen in Figures 1 and 2, mean change scores were lowest when beliefs were associated with higher affect than knowledge scores, as was the case for belief in creationism, an afterlife, and a disbelief in evolution. This relationship was tested by correlating participants' knowledge-affect difference scores with their willingness to change their beliefs on the five topics. All five bi-variate correlations were significant at $p < .05$, and the coefficients ranged between .30 and .40. Thus, the more a participant's belief was based in knowledge relative to affect, the more willing they were to change that belief in the face of conflicting evidence.

In sum, many participants reported that their beliefs were based more on affect than on any relation to existing knowledge. Also, there was significant variation among beliefs in terms of their knowledge versus affective bases. Lastly, participants claimed they would be less willing to change affect-based beliefs than knowledge-based beliefs if presented with sound belief conflicting evidence.

Table 3. Correlations between knowledge-affect difference scores and belief, on all five topics.

Creationism	Afterlife	ESP	Opposites	Evolution
-.68*	-.38*	-.42*	-.35*	.73*

*$p < .05$.

Discussion

These preliminary results support our hypothesis that beliefs and knowledge are related but distinct constructs. People will not only report that some of their beliefs are held on affective grounds, they will even specifically reject knowledge based reasons as the bases for some of their beliefs. In addition, these self-reported reasons for belief predicted participants willingness to change those beliefs.

These data are only a first attempt to examine this issue. It is a difficult task to assess the true bases for individual beliefs. If belief formation and maintenance are relatively deliberate mental enterprises then it is reasonable to assume that people would be able to accurately report the relationships between their beliefs and relevant knowledge and affective goals. The validity of these self-reports is aided by the fact that

participants' self-generated reasons matched their endorsement of the knowledge and affective reasons that we provided.

It should be noted that self-reported willingness to change a belief was not intended to be an actual measure of belief revision. This point is made obvious by the fact that no belief conflicting information was ever provided to participants. However, we argue that belief revision is highly subject to motivational influence and that epistemological values are integral to people's motivation to update beliefs with any newly acquired knowledge. Self-reported willingness to change a belief reflects belief-specific epistemological values that should affect the motivations relevant to belief revision. Thus, it is noteworthy that participants reported being rather unwilling to change their affect-based beliefs, even if presented with sound conflicting evidence, but relatively willing to change knowledge-based beliefs.

Implications and Future Research

The obvious next step is to see whether these reported differences in belief bases predict how well people comprehend new information in conflict with their beliefs. The distinction between knowledge and beliefs requires that outcome measures be tailored to assess change in one or the other. If conceptual understanding (knowledge change) is being assessed, then participants must be clear that their task is to demonstrate their understanding of the new concepts, and not to report their current point of view.

A similar concern arises when the outcome of interest is belief revision or conceptual replacement. Dependent measures must show that people are spontaneously employing new concepts in their thinking, not merely adapting their thinking to the expectations of experimental or educational settings. Being explicit in our discussions and methodologies about beliefs versus knowledge may reveal where the real disagreements are in the area of conceptual change and perhaps reveal that there is less disagreement than it seems.

Previous accounts of conceptual change have assumed a uniformly negative influence of prior beliefs on conceptual change. We argue that the knowledge versus affective basis of prior beliefs may be an important determinant of whether conflicting concepts are accurately understood. We also contend that the coherence competition accounts of belief revision are too simplistic, given the existence of affect-based beliefs. A conceptual-coherence theory of belief revision only makes sense for the sub-set of beliefs that are initially based on their coherence with current knowledge.

Previous theories claim that conceptual change is in the direction of greater coherence. Thus, incoherent prior beliefs should be more likely to change than coherent beliefs. The present theory suggests several reasons to expect just the opposite in some circumstances. Affect-based beliefs by virtue of their lack of coherence with the conceptual framework might be immune to threats posed by conflicting information. Any new information is likely to be distorted, and if it is accurately comprehended, it will have little influence on an affect-based belief.

The present theory predicts that emotional beliefs not derived from relevant knowledge are the least likely to change in the face of conflicting information. This prediction should make intuitive sense to anyone who has ever had a dinner-time discussion about politics or religion. Scientific ideas may change slowly, but the informal observation that they seem to change quicker than non-scientific ideas may be an indication that the scientific enterprise generally does adhere to the principles of forming ideas based on knowledge and coherent argument.

References

Chi, M. T. H. (1992). Conceptual change within and across ontological categories: Examples from learning and discovery in science. In R. N. Giere (Ed.), Minnesota studies in the philosophy of science: Vol. XV, Cognitive models of science (pp. 129-186). Minneapolis: University of Minnesota Press.

Chinn, C. A., & Brewer, W. F. (2000). Knowledge change in response to data in science, religion, and magic. In K. S. Rosengren, C, N. Johnson, and P. L. Harris (Eds.), Imagining the impossible: Magical, scientific, and religious thinking in children (pp. 334-371). Cambridge, U. K.: Cambridge University Press.

diSessa, A. A. (2000). A complex adaptive systems view of conceptual change. In L. R. Gleitman and A. K. Joshi (Eds.), Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society (pp. 8-9). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Dole, J. A., & Sinatra, G. M. (1998). Reconceptualizing change in the cognitive construction of knowledge. Educational Psychologist, 33, 109-128.

Eagly, A. H., & Chaiken, S. (1993). The psychology of attitudes. London: Harcourt Brace Publishers.

Guzzetti, B. J., Snyder, T. E., Glass, G. V., & Gamas, W. S. (1993). Promoting conceptual change in science: A comparative meta-analysis of instructional interventions from reading education and science education. Reading Research Quarterly, 28, 117-159.

Ohlsson, S., & Lehtinen, E. (1997). Abstraction and the acquisition of complex ideas. International Journal of Educational Research, 27, 37-48.

Quine, W. V., & Ullian, J. S. (1970). The web of belief. New York: Random House.

Thagard, P. (1992). Conceptual revolutions. Princeton, NJ: Princeton University Press.

Randomness and Coincidences: Reconciling Intuition and Probability Theory

Thomas L. Griffiths & Joshua B. Tenenbaum
Department of Psychology
Stanford University
Stanford, CA 94305-2130 USA
{gruffydd,jbt}@psych.stanford.edu

Abstract

We argue that the apparent inconsistency between people's intuitions about chance and the normative predictions of probability theory, as expressed in judgments about randomness and coincidences, can be resolved by focussing on the evidence observations provide about the processes that generated them rather than their likelihood. This argument is supported by probabilistic modeling of sequence and number production, together with two experiments that examine judgments about coincidences.

People are notoriously inaccurate in their judgments about randomness, such as whether a sequence of heads and tails like HHTHTTTH is more random than the sequence HHHHHHHH. Intuitively, the former sequence seems more random, but both sequences are equally likely to be produced by a random generating process that chooses H or T with equal probability, such as a fair coin. This kind of question is often used to illustrate how our intuitions about chance deviate from the normative standards set by probability theory. Our intuitions about coincidental events, which seem to be defined by their improbability, have faced similar criticism from statisticians (eg. Diaconis & Mosteller, 1989).

The apparent inconsistency between our intuitions about chance and the formal structure of probability theory has provoked attention from philosophers and mathematicians, as well as psychologists. As a result, a number of definitions of randomness exist in both the mathematical (eg. Chaitin, 2001; Kac, 1983; Li & Vitanyi, 1997) and the psychological (eg. Falk, 1981; Lopes, 1982) literature. These definitions vary in how well they satisfy our intuitions, and can be hard to reconcile with probability theory. In this paper, we will argue that there is a natural relationship between people's intuitions about chance and the normative standards of probability theory. Traditional criticism of people's intuitions about chance has focused on the fact that people are poor estimators of the likelihood of events being produced by a particular generating process. The models we present turn this question around, asking how much more likely a set of events makes a particular generating process. This question may be far more useful in natural inference situations, where it is often more important to reason diagnostically than predictively, attempting to infer the structure of our world from the data we observe.

Randomness

Reichenbach (1934/1949) is credited with having first suggested that mathematical novices will be unable to produce random sequences, instead showing a tendency to overestimate the frequency with which outcomes alternate. Subsequent research has provided support for this claim (reviewed in Bar-Hillel & Wagenaar, 1991; Tune, 1964; Wagenaar, 1972), with both sequences of numbers (eg. Budescu, 1987; Rabinowitz, Dunlap, Grant, & Campione, 1989) and two-dimensional black and white grids (Falk, 1981). In producing binary sequences, people alternate with a probability of approximately 0.6, rather than the 0.5 that is seen in sequences produced by a random generating process. This preference for alternation results in subjectively random sequences containing less runs – such as an interrupted series of heads in a set of coin flips – than might be expected by chance (Lopes, 1982).

Theories of subjective randomness

A number of theories have been proposed to account for the accuracy of Reichenbach's conjecture. These theories have included postulating that people develop a concept of randomness that differs from the true definition of the term (eg. Budescu, 1987; Falk, 1981; Skinner, 1942), and that limited short-term memory might contribute to people's responses (Baddeley, 1966; Kareev, 1992; 1995; Wiegersma, 1982). Most recently, Falk and Konold (1997) suggested that the concept of randomness can be connected to the subjective complexity of a sequence, characterized by the difficulty of specifying a rule by which a sequence can be generated. This idea is related to a notion of complexity based on description length (Li & Vitanyi, 1997), and has been considered elsewhere in psychology (Chater, 1996).

The account of randomness that has had the strongest influence upon the wider literature of cognitive psychology is Kahneman and Tversky's (1972) suggestion that people may be attempting to produce sequences that are "representative" of the output of a random generating process. For sequences, this means that the number of elements of each type appearing in the sequence should correspond to the overall probability with which these elements occur. Random sequences should also maintain local representativeness, such that subsequences demonstrate the appropriate probabilities.

Formalizing representativeness

A major challenge for a theory of randomness based upon representativeness is to express exactly what it means for an outcome to be representative of a random generating process. One interpretation of this statement is that the outcome provides evidence for having been produced by a random generating process. This interpretation has the advantage of submitting easily to formalization in the language of probability theory.

If we are considering two candidate processes by which an outcome could be generated – one random, and one containing systematic regularities – the total evidence in favor of the random generating process can be assessed by the logarithm of the ratio of the probabilities of these processes

$$\log \frac{P(\text{random}|x)}{P(\text{regular}|x)}, \tag{1}$$

where $P(\text{random}|x)$ and $P(\text{regular}|x)$ are the probabilities of a random and a regular generating process respectively, given the outcome x.

This quantity can be computed using the odds form of Bayes' rule

$$\frac{P(\text{random}|x)}{P(\text{regular}|x)} = \frac{P(x|\text{random})}{P(x|\text{regular})} \frac{P(\text{random})}{P(\text{regular})}, \tag{2}$$

in which the term on the left-hand side of the equation is called the posterior odds, and the first and second terms on the right-hand side are called the likelihood ratio and prior odds, respectively. Of the latter two terms, the specific outcome x influences only the likelihood ratio. Thus the contribution of x to the evidence in favour of a random generating process can be measured by the logarithm of the likelihood ratio,

$$\text{random}(x) = \log \frac{P(x|\text{random})}{P(x|\text{regular})}. \tag{3}$$

This method of assessing the weight of evidence for a particular hypothesis provided by an observation is often used in Bayesian statistics, and the log likelihood-ratio given above is called a Bayes factor (Kass & Raftery, 1995). The Bayes factor for a set of independent observations will be the sum of their individual Bayes factors, and the expression has a clear information theoretic interpretation (Good, 1979). The above expression is also closely connected to the notion of minimum description length, connecting this approach to randomness with the ideas of Falk and Konold (1997) and Chater (1996).

Defining regularity

Evaluating the evidence that a particular outcome provides for a random generating process requires computing two probabilities: $P(x|\text{random})$ and $P(x|\text{regular})$. The first of these probabilities follows from the definition of the random generating process. For example, $P(\texttt{HHTHTTTH}|\text{random})$ is $(\frac{1}{2})^8$, as it would be for any sequence of the same length. However, computing $P(x|\text{regular})$ requires specifying the probability of the observed outcome resulting from a generating process that involves regularities. While this probability is hard to define, it is in general easy to compute $P(x|h_i)$, where h_i might be some hypothesised regularity. In the case of sequences of heads and tails, for instance, h_i might correspond to a particular probability of observing heads, $P(\texttt{H}) = p$. In this case $P(\texttt{HHTHTTTH}|h_i)$ is $p^4(1-p)^4$. Using the calculus of probability, we can obtain $P(x|\text{regular})$ by summing over a set of hypothesized regularities, $\mathcal{H}$,

$$P(x|\text{regular}) \quad = \quad \sum_{h_i \in H} P(x|h_i)P(h_i|\text{regular}) \tag{4}$$

where $P(h_i|\text{regular})$ is a prior probability on h_i. In all applications discussed in this paper, we make the simplifying assumption that $P(h_i|\text{regular})$ is uniform over all $h_i \in \mathcal{H}$. However, we stress that this assumption is not necessary for the models we create, and the prior may in fact differ from uniformity in some realistic judgment contexts.

Random sequences

For the case of binary sequences, such as those that might be produced by flipping a coin, possible regularities can be divided into two classes. One class assumes that flips are independent, and the regularities it contains are assertions about the value of $P(\texttt{H})$. The second class includes regularities that make reference to properties of subsequences containing more than a single element, such as alternation, runs, and symmetries. Since this second class is less well defined, it is instructive to examine the account that can be obtained just by using the first class of regularities.

Taking $\mathcal{H}$ to be all values of $p = P(\texttt{H}) \in [0,1]$, we have $P(H,T|\text{random}) = (\frac{1}{2})^{H+T}$ and $P(H,T|\text{regular}) = \int_0^1 p^H(1-p)^T dp$, where H,T are the sufficient statistics of a particular sequence containing H heads and T tails. Completing the integral, it follows from (3) that

$$\text{random}(H,T) = \log(\tbinom{H+T}{H}) + f(H+T), \tag{5}$$

where $f(H+T)$ is $-\log 2^{H+T} - \log(H+T+1)$, a fixed function of the total number of flips in the sequence. This result has a number of appealing properties. Firstly, it is maximized when $H = T$, which is consistent with Kahneman and Tversky's (1972) original description of the representativeness of random sequences. Secondly, the ratio involved essentially measures the size of the set of sequences sharing the same number of heads and tails. A sequence like HHHHHHHH is unique in its composition, whereas HHTHTTTH has a composition much more commonly obtained by flipping a coin eight times.

The Zenith radio data

Having defined a framework for analyzing the subjective randomness of sequences, we have the opportunity to develop a specific model. One classic data set concerning

the production of random sequences is the Zenith radio data. These data were obtained as a result of an attempt by the Zenith corporation to test the hypothesis that people are sensitive to psychic transmissions. On several occasions in 1937, a radio program took place during which a group of psychics would transmit a randomly generated binary sequence to the receptive minds of their listeners. The listeners were asked to write down the sequence that they received, one element at a time. The binary choices included heads and tails, light and dark, black and white, and several symbols commonly used in tests of psychic abilities, and all sequences contained a total of five symbols. Listeners then mailed in their responses, which were analyzed. These responses demonstrated strong preferences for particular sequences, but there was no systematic effect of the actual sequence that was transmitted (Goodfellow, 1938). The data are thus a rich source of information about response preferences for random sequences. The relative frequencies of the different sequences, collapsed over choice of first symbol, are shown in the upper panel of Figure 1.

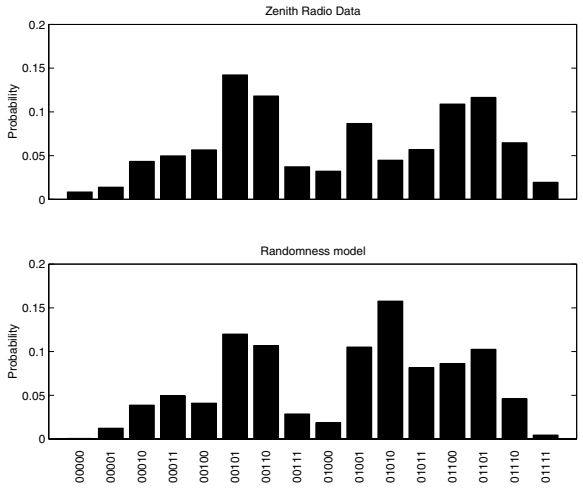


Figure 1: The upper panel shows the original Zenith radio data, representing the responses of 20,099 participants, from Goodfellow (1938). The lower panel shows the predictions of the randomness model. Sequences are collapsed over the initial choice, represented by 0.

Modeling random sequence production

One of the most important characteristics of the Zenith radio data is that people's responses were produced sequentially. In producing each element of the sequence, people had knowledge of the previous elements. Kahneman and Tversky (1972) suggested that in producing such sequences, people pay attention to the local representativeness of their choices – the representativeness of each subsequence.

To capture this idea, we define L_k to be the local representativeness of choosing `H` as the kth response – the extent to which `H` results in a more random outcome than `T`, assessed over the subsequences starting one step back, two steps back, and so forth,

$$L_k = \sum_{i=1}^{k-1} \text{random}(H_i + 1, T_i) - \text{random}(H_i, T_i + 1), \quad (6)$$

where the H_i, T_i are the tallies of heads and tails counting back i steps in the sequence. We can then convert this quantity into a probability using a logistic function, to give a probability distribution for the kth response, R_k:

$$\begin{aligned} P(R_k = \texttt{H}) &= \frac{1}{1 + e^{-\lambda L_k}} \\ &= \frac{1}{1 + \prod_{i=1}^{k-1} \left(\frac{T_i+1}{H_i+1}\right)^{-\lambda}} \quad (7) \\ &= \frac{\prod_{i=1}^{k-1} (T_i+1)^{\lambda}}{\prod_{i=1}^{k-1} (T_i+1)^{\lambda} + \prod_{i=1}^{k-1} (H_i+1)^{\lambda}}. \quad (8) \end{aligned}$$

The λ parameter scales the effect that L_k has on the resulting probability. The probability of the sequence as a whole is then the product of the probabilities of the R_k, and the result defines a probability distribution over the set of binary sequences of length k. This distribution is shown in the lower panel of Figure 1 for $k = 5$.

This simple model provides a remarkably good account of the response preferences people demonstrated in the Zenith radio experiment. There is one clear discrepancy: the model predicts that the sequence `01010`, equivalent to `HTHTH` or `THTHT`, should occur far more often than in the data. We can explain people's avoidance of this sequence by the fact that alternation itself forms a regularity, which could easily be introduced into the hypothesis space. More striking is the account the model gives of the different frequencies of sequences with less apparent regularities, such as `00001` and `00010`. Excluding the discrepant data point, the model gives a parameter-free ordinal correlation $r_s = 0.97$, and with $\lambda = 0.6$ has a linear correlation $r = 0.95$. Interestingly, the model predicts alternation, for sequences that are otherwise equally representative, with a probability of $\frac{1}{1+2^{-\lambda}}$. With the value of λ used in fitting the Zenith radio data, the resulting predicted probability of alternation is 0.6, consistent with previous findings (eg. Falk, 1981).

Pick a number

Research on subjective randomness has focused almost exclusively on sequences, but sequences are not the only stimuli that excite our intuitions about chance. In particular, random numbers loom larger in life than in the literature, although there have been a few studies that have investigated response preferences for numbers between 0 and 9. Kubovy and Psotka (1976) reported the frequency with which people produce numbers between 0 and 9 when asked to pick a number, aggregated across several studies. These results are shown in the upper panel of Figure 2. People showed a clear preference for

the number 7, which Kubovy and Psotka (1976, p. 294) explained with reference to the properties of the numbers involved – for example, 6 is even, and a multiple of 3, but it is harder to find properties of 7. This explanation is suggestive of the kinds of regular generating processes that could be involved in producing numbers. Shepard and Arabie (1979) found that similarity judgments about numbers could be captured by properties like those described by Kubovy and Psotka (1976), such as being even numbers, powers of 2, or occupying special positions such as endpoints.

Taking the arithmetic properties of numbers to constitute hypothetical regularities, we can specify the quantities necessary to compute random(x). Our h_i are sets of numbers that share some property, such as the set of even numbers between 0 and 9. For any h_i, we define $P(x|h_i) = \frac{1}{|h_i|}$ for $x \in h_i$ and 0 otherwise, where $|h_i|$ is the size of the set. This means that observations generated from a regularity are uniformly sampled from that regularity. Setting $P(h_i|\text{regular})$ to give equal weight to all h_i, we can compute $P(x|\text{regular})$.

This model can be applied to the data of Kubovy and Psotka (1976). Since there are ten possible responses, we have $P(x|\text{random}) = \frac{1}{10}$. Taking hypothetical regularities of multiples of 2 ($\{0,2,4,6,8\}$), multiples of 3 ($\{3,6,9\}$), multiples of 5 ($\{0,5\}$), powers of 2 ($\{2,4,8\}$), and endpoints ($\{0\},\{1\},\{9\}$), we obtain the values of random(x) shown in the lower panel of Figure 2. Randomness also needs to be included in $\mathcal{H}$ so that random(x) is defined when x is not in any other regularity. Its inclusion is analogous to the incorporation of a noise process, and is in fact formally identical in this case. The order of the model predictions is a parameter free result, and gives the ordinal correlation $r_s = 0.99$. Applying a single parameter power transformation to the predictions, $y' = (y - \min(y))^{0.98}$, gives $r = 0.95$.

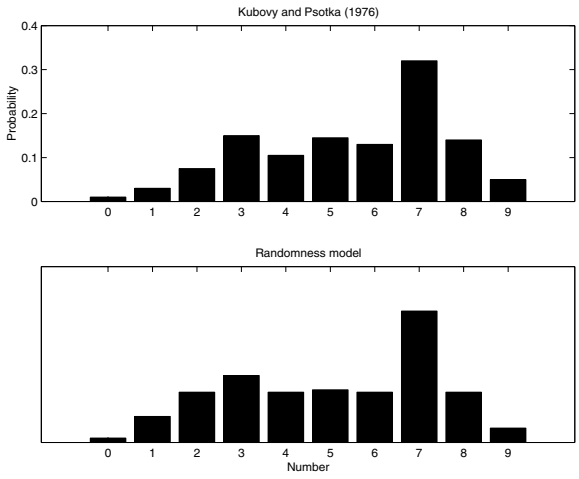


Figure 2: The upper panel shows number production data from Kubovy and Psotka (1976), taken from 1,770 participants choosing numbers between 0 and 9. The lower panel shows the transformed predictions of the randomness model.

Coincidences

The surprising frequency with which unlikely events tend to occur has drawn attention from a number of psychologists and statisticians. Diaconis and Mosteller (1989), in their analysis of such phenomena, define a coincidence as '...a surprising concurrence of events, perceived as meaningfully related, with no apparent causal connection' (p. 853). They go on to suggest that the "surprising" frequency of these events is due to the flexibility that we allow in identifying meaningful relationships. Together with the fact that everyday life provides a vast number of opportunities for coincidences to occur, our willingness to tolerate near misses and to consider each of a number of possible concurrences meaningful contributes to explaining the frequency with which coincidences occur. Diaconis and Mosteller suggested that the surprise that people show at the solution to the Birthday Problem – the fact that only 23 people are required to give a 50% chance of two people sharing the same birthday – suggests that similar neglect of combinatorial growth contributes to the underestimation of the likelihood of coincidences. Psychological research addressing coincidences seems consistent with this view, suggesting that selective memory (Hintzman, Asher, & Stern, 1978) and preferential weighting of first-hand experiences (Falk & MacGregor, 1983) might facilitate the under-estimation of the probability of events.

Not just likelihood...

The above analyses reflect the same bias that made it difficult to construct a probabilistic account of randomness: the notion that people's judgments reflect the likelihood of particular outcomes. Subjectively, coincidences are events that seem unlikely, and are hence surprising when they occur. However, just as with random sequences, sets of events that are equally likely to be produced by a random generating process differ in the degree to which they seem to be coincidences. Following Diaconis and Mosteller's suggestion that the Birthday Problem provides a domain for the investigation of coincidences, consider the kinds of coincidences formed by sets of birthdays. If we meet four people and find out that their birthdays are October 4, October 4, October 4, and October 4, this is a much bigger coincidence than if the same people have birthdays May 14, July 8, August 21, and December 25, despite the fact that these sets of birthdays are equally likely to be observed by chance. The way that these sets of birthdays differ is that one of them contains an obvious regularity: all four birthdays occur on the same day.

Modeling coincidences

Just as sequences differ in the amount of evidence they provide for having been produced by a random generating process, sets of birthdays differ in how much evi-

dence they provide for having been produced by a process that contains regularities. We argue that the amount of evidence that an event provides for a regular generating process will correspond to how big a coincidence it seems, and that this can be computed in the same way as for randomness,

$$\text{coincidence}(x) = \log \frac{P(x|\text{regular})}{P(x|\text{random})}. \qquad (9)$$

To apply this model we have to define the regularities $\mathcal{H}$. For birthdays, these regularities should correspond to relationships that can exist among dates. Our model of coincidences used a set of regularities that reflected proximity in date (from 1 to 30 days), belonging to the same calendar month, and having the same calendar date (eg. January 17, March 17, September 17, December 17). We also assumed that each year consists of 12 months of 30 days each. Thus, for a set of n birthdays, $X = \{x_1, \ldots, x_n\}$, we have $P(X|\text{random}) = (\frac{1}{360})^n$. In defining $P(X|\text{regular})$, we want to respect the fact that regularities among birthdays are still striking even when they are embedded in noise – for instance, February 2, March 26, April 3, June 12, June 12, June 12, June 12, November 22 still provides strong evidence for a regularity in the generating process. To allow the model to tolerate noisy regularities, we can introduce a noise term α into $P(X|h_i)$. The probability calculus lets us integrate out unwanted parameters, so the introduction of a noise process need not result in adding a numerical free parameter to the model. In particular, $P(X|h_i) = \int_0^1 P(X|\alpha, h_i)P(\alpha|h_i)d\alpha$. Assuming that the dates we observe are independent, we have $P(X|\alpha, h_i) = \prod_{x_j \in X} P(x_j|\alpha, h_i)$, and, taking a uniform prior on α, $P(X|h_i)$ is simply $\int_0^1 \prod_{x_j \in X} P(x_j|\alpha, h_i)d\alpha$, where

$$P(x_j|\alpha, h_i) = \begin{cases} \frac{\alpha}{360} + (1-\alpha)\frac{1}{|h_i|} & x_j \in h_i \\ \frac{\alpha}{360} & x_j \notin h_i \end{cases}. \qquad (10)$$

This corresponds to dates being sampled uniformly from the entire year with probability α, and uniformly from the regularity with probability $(1-\alpha)$. The resulting $P(X|h_i)$ can then be substituted into (4), and taking a uniform distribution for $P(h_i|\text{regular})$ gives $P(X|\text{regular})$.

How big a coincidence?

The model outlined above makes strong predictions about the degree to which different sets of birthdays should be judged to constitute coincidences. We conducted a simple experiment to examine these predictions. The participants were 93 undergraduates from Stanford University, participating for partial course credit. Fourteen potential relationships between birthdays were examined, using two sets of dates. Each participant saw one set of dates, in a random order. The dates reflected: 2, 4, 6, and 8 apparently unrelated birthdays, 2 birthdays on the same day, 2 birthdays in 2 days across a month boundary, 4 birthdays on the same day, 4 birthdays in one week across a month boundary, 4 birthdays in the same calendar month, 4 birthdays with the same calendar dates, and 2 same day, 4 same day, and 4 same date with an additional 4 unrelated birthdays, as well as 4 same week with an additional 2 unrelated birthdays. These dates were delivered in a questionnaire. Each participant was instructed to rate how big a coincidence each set of dates was, using a scale in which 1 denoted no coincidence and 10 denoted a very big coincidence.

The results of the experiment and the model predictions are shown in the top and middle panels of Figure 3 respectively . Again, the ordinal predictions of the model are parameter free, with $r_s = 0.94$. Applying the transformation $y' = (y - \min(y))^{0.48}$, gives $r = 0.95$. The main discrepancies between the model and the data are the four birthdays that occur in the same calendar month, and the ordering of the random dates. The former could be addressed by increasing the prior probability given to the regularity of being in the same calendar month – clearly this was given greater weight by the participants than by the model. Explaining the increase in the judged coincidence with larger sets of unrelated dates is more difficult, but may be a result of opportunistic coincidences: as more dates are provided, participants have more opportunities to identify complex regularities or find dates of personal relevance. This process can be incorporated into the model, at the cost of greater complexity.

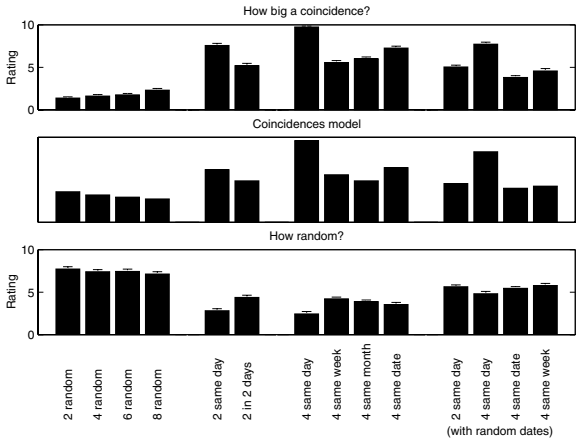


Figure 3: The top panel shows the judged extent of coincidence for each set of dates. The middle panel is the predictions of the coincidences model, subjected to a transformation described in the text. The bottom panel shows randomness judgments for the same stimuli.

Relating randomness and coincidences

Judgments of randomness and coincidences both reflect the evidence that a set of observations provides for having been produced by a particular generating process. Events that provide good evidence for a random generating process are viewed as random, while events that provide evidence for a generating process incorporating some regularity seem like coincidences. By examining

(3) and (9), we see that these phenomena are formally identified as inversely related: coincidences are events that deviate from our notions of randomness.

We conducted a further experiment to see if this relationship was borne out in people's judgments. Participants were 120 undergraduates from Stanford University, participating for partial course credit. The dates were the same as those used previously, and delivered in similar format. Each participant was instructed to rate how random each set of dates was, using a scale in which 1 denoted not at all random and 10 denoted very random.

The results of this experiment are shown in the bottom panel of Figure 3. The correlation between the randomness judgments and the coincidence judgments is $r = -0.94$, consistent with the hypothesis that randomness and coincidences are inversely linearly related. The main discrepancy between the two data sets is that the addition of unrelated dates seems to affect randomness judgments more than coincidence judgments.

Conclusion

The models we have discussed in this paper provide a connection between people's intuitions about chance, expressed in judgments about randomness and coincidences, and the formal structure of probability theory. This connection depends upon changing the way we model questions about probability. Rather than considering the likelihood of events being produced by a particular generating process, our models address the question of how much more likely a set of events makes a particular generating process. This is a structural inference, drawing conclusions about the world from observed data. Framed in this way, people's judgments are revealed to accurately approximate the statistical evidence that observations provide for having been produced by a particular generating process. The apparent inaccuracy of our intuitions may thus be a result of considering normative theories based upon the likelihood of events rather than the evidence they provide for a structural inference.

References

Baddeley, A. D. (1966). The capacity of generating information by randomization. *Quarterly Journal of Experimental Psychology*, 18:119–129.

Bar-Hillel, M. and Wagenaar, W. A. (1991). The perception of randomness. *Advances in applied mathematics*, 12:428–454.

Budescu, D. V. (1987). A Markov model for generation of random binary sequences. *Journal of Experimental Psychology: Human perception and performance*, 12:25–39.

Chaitin, G. J. (2001). *Exploring randomness*. Springer Verlag, London.

Chater, N. (1996). Reconciling simplicity and likelihood principlesin perceptual organization. *Psychological Review*, 103:566–581.

Diaconis, P. and Mosteller, F. (1989). Methods for studying coincidences. *Journal of the American Statistical Association*, 84:853–861.

Falk, R. (1981). The perception of randomness. In *Proceedings of the fifth international conference for the psychology of mathematics education*, volume 1, pages 222–229, Grenoble, France. Laboratoire IMAG.

Falk, R. and Konold, C. (1997). Making sense of randomness: Implicit encoding as a bias for judgment. *Psychological Review*, 104:301–318.

Falk, R. and MacGregor, D. (1983). The surprisingness of coincidences. In Humphreys, P., Svenson, O., and Vari, A., editors, *Analysing and aiding decision processes*, pages 489–502. North-Holland, New York.

Good, I. J. (1979). A. M. Turing's statistical work in World War II. *Biometrika*, 66:393–396.

Goodfellow, L. D. (1938). A psychological interpretation of the results of the Zenith radio experiments in telepathy. *Journal of Experimental Psychology*, 23:601–632.

Hintzman, D. L., Asher, S. J., and Stern, L. D. (1978). Incidental retrieval and memory for coincidences. In Gruneberg, M. M., Morris, P. E., and Sykes, R. N., editors, *Practical aspects of memory*, pages 61–68. Academic Press, New York.

Kac, M. (1983). What is random? *American Scientist*, 71:405–406.

Kahneman, D. and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3:430–454.

Kareev, Y. (1992). Not that bad after all: Generation of random sequences. *Journal of Experimental Psychology: Human Perception and Performance*, 18:1189–1194.

Kareev, Y. (1995). Through a narrow window: Working memory capacity and the detection of covariation. *Cognition*, 56:263–269.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kubovy, M. and Psotka, J. (1976). The predominance of seven and the apparent spontaneity of numerical choices. *Journal of Experimental Psychology: Human Perception and Performance*, 2:291–294.

Li, M. and Vitanyi, P. (1997). *An introduction to Kolmogorov complexity and its applications*. Springer Verlag, London.

Lopes, L. (1982). Doing the impossible: A note on induction and the experience of randomness. *Journal of Experimental Psycholgoy*, 8:626–636.

Rabinowitz, F. M., Dunlap, W. P., Grant, M. J., and Campione, J. C. (1989). The rules used by children and adults to generate random numbers. *Journal of Mathematical Psychology*, 33:227–287.

Reichenbach, H. (1934/1949). *The theory of probability*. University of California Press, Berkeley.

Shepard, R. and Arabie, P. (1979). Additive clutering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, 86:87–123.

Skinner, B. F. (1942). The processes involved in the repeated guessing of alternatives. *Journal of Experimental Psychology*, 39:322–326.

Tune, G. S. (1964). Response preferences: A review of some relevant literature. *Psychological Bulletin*, 61:286–302.

Wagenaar, W. A. (1972). Generation of random sequences by human subjects: A critical survey of literature. *Psychological Bulletin*, 77:65–72.

Wiegersma, S. (1982). Can repetition avoidance in randomization be explained by randomness concepts? *Psychological Research*, 44:189–198.

Acknowledgments

This work was supported by a Hackett studentship to the first author and funds from Mitsubishi Electric Research Laboratories. The authors thank Persi Diaconis for inspiration and conversation, Tania Lombrozo, Craig McKenzie and an anonymous reviewer for helpful comments, and Kevin Lortie for finding the leak.

Judging the Probability of Representative and Unrepresentative Unpackings

Constantinos Hadjichristidis (constantinos.hadjichristidis@durham.ac.uk)
Department of Psychology, University of Durham
Durham, DH1 3LE, UK

Steven A. Sloman (steven_sloman@brown.edu)
Department of Cognitive & Linguistic Sciences, Brown University
Box 1978, Providence, RI 02912, USA

Edward J. Wisniewski (edw@uncg.edu)
Department of Psychology, University of North Carolina at Greensboro
Box 26164, Greensboro, NC 27402-6164, USA

Abstract

The hypothesis that category descriptions are interpreted narrowly, in terms of representative instances, is examined by comparing probability judgments for packed descriptions of events to judgments for coextensional unpacked descriptions. The representativeness of the unpacked instance was varied along with the type of unpacking (direct vs. priming). In contrast to the prediction of Support Theory (Tversky & Koehler, 1994), we found no evidence that unpacking has a nonnegative effect on probability judgments (subadditivity). Instead, we found a negative effect with low representative direct unpackings (superadditivity). Our data suggest that probability judgments are proportional to the typicality of instances in the description.

Introduction

People frequently assess the probability of uncertain events such as the chance of rain or the success rate of a medical treatment. Such probability assessments are important because they determine not only whether to plan a barbecue, but also whether or not to have surgery. Judgments concerning rain or effectiveness of a treatment are categorical in the sense that the event being judged could be instantiated in many ways. We consider whether the typicality of the instances used in a categorical description affects the judged probability of corresponding events.

Normative models of probability judgment assume *description invariance*: the probability of an event does not depend on how the event is described. This assumption is descriptively invalid (Tversky & Kahneman, 1986). People give lower probability ratings for the packed hypothesis "Death from homicide, rather than accidental death" than for the coextensional unpacked hypothesis "Death from homicide by an acquaintance or by a stranger, rather than accidental death" (Rottenstreich & Tversky, 1997).

Support for Support Theory?

To accommodate this fact, Tversky and Koehler (1994) proposed a descriptive theory of probability judgment, Support Theory, that suggests that subjective probabilities are assigned not to events, but to descriptions of events or *hypotheses*. Probability judgments are hypothesized to be mediated by evaluations of evidence for and against a hypothesis. Specifically, the judged probability of a hypothesis H rather than an alternative hypothesis A is given by:

$$(1) \quad \text{Judged } P(H, A) = s(H) / [s(H) + s(A)]$$

where s(X) is a global measure of support for hypothesis X. It is a sum of the (weighted) support of all representative instances of X that are available to the judge at the time of evaluation.

A key assumption of Support Theory is that exhaustively unpacking a hypothesis H into mutually exclusive and exhaustive sub-hypotheses ($H_1 \vee \ldots \vee H_n$) can increase the support for H:

$$(2) \quad s(H) \leq s(H_1 \vee \ldots \vee H_n)$$

This assumption is motivated on the grounds that unpacking may bring additional instances to mind, or increase the salience of the unpacked instances. Either or both of these effects would increase the perceived support for a hypothesis.

Taken together, Support Theory's assumptions predict implicit subadditivity: The judged probability of an implicit (or packed) hypothesis H is no greater than the judged probability of a coextensional unpacked hypothesis. Implicit subadditivity has been observed several times (see Rottenstreich & Tversky, 1997).

However, not all the data are so supportive. Hadjichristidis et al. (1999) showed that selectively unpacking hypotheses into components that enjoy low levels of support results in the opposite phenomenon, implicit superadditivity. To illustrate, students gave higher probability estimates for the packed hypothesis "death from a natural cause" than for its coextensional unpacked counterpart "death from asthma, the flu, or some other natural cause." In a series of follow-up studies we have consistently found implicit superadditivity with novel categories unpacked with atypical instances. We have consistently failed to find implicit subadditivity, even when events were unpacked using representative instances, instances that enjoy high levels of support. In sum, contrary to support theory's predictions, these data suggest that unpacking does not always increase subjective probability judgments.

The Supported Theory

A parsimonious interpretation of our data is based on Support Theory's own assumption that people interpret category-based hypotheses narrowly, in terms of representative instances. Unpacking unrepresentative instances induces superadditivity by making instances of very low support part of what is judged. Unpacking representative instances leaves probability judgments unaffected because packed categories are interpreted in terms of representative instances.

Unlike Support Theory, we suggest that unpacking can decrease support. According to the present proposal, people assess the likelihood of a category-based hypothesis by thinking about instances in which the event is expected to occur (i.e., by bringing to mind representative instances, instances enjoying high levels of support). This dovetails with Kahneman and Miller's (1986) proposal that norms—contrast events for judgments of surprise, blame, etc.—are constructed according to the availability and representativeness of exemplars. Our proposal is that the determinants of exemplar retrieval control not only how contrast events are conceived, but how focal ones are too. Moreover, the mere availability in memory of an instance is not sufficient to change judgments of likelihood, the instance must be one of the objects of judgment.

Study

The current study tests our hypothesis by crossing representativeness (high- vs. low-representative instances) with type of unpacking (direct vs. priming) in a between-participants design. A separate group of participants was asked to provide estimates for corresponding packed hypotheses. The dependent measure was subjective probability judgment. Table 1 gives one illustration from each of the five experimental conditions.

Table 1: An example stimulus from each of the five conditions. The sentence in bold-faced letters is a description that preceded evaluations in all conditions.

	Sarah is a very energetic and happy eight year old who loves playing with her stuffed animals.
Packed	How likely is it that Sarah hates some types of pets (as opposed to loving all pets)? _____
Direct High Rep	How likely is it that Sarah hates tarantulas or some other types of pets (as opposed to loving all pets)? _____
Direct Low Rep	How likely is it that Sarah hates horses or some other types of pets (as opposed to loving all pets)? _____
Priming High Rep	Same as packed but prior to making the judgment primed with a list of words including "tarantulas"
Priming Low Rep	Same as packed but prior to making the judgment primed with a list of words including "horses"

Direct unpacking refers to a conventional unpacking manipulation. Participants in the direct unpacking conditions were asked to judge categories from which one of their instances had been unpacked. Based on previous findings, we expected to find a negative effect of unpacking unrepresentative instances (i.e., implicit superadditivity), and no effect of unpacking representative instances.

Participants in the priming unpacked conditions were asked to judge packed hypotheses after being primed with either representative or unrepresentative instances. Priming consisted of asking participants to study the instances for 1 min. for a later memory test. We reasoned that priming would make the critical instances highly available in memory at the time of judgment without specifically making them the objects of judgment. If merely making an instance available in memory increases the likelihood that it will be considered during judgment of a category that is superordinate to it, then superadditivity should be observed in the Priming Low Representativeness condition (i.e., probability judgments should be higher in the Packed condition than the Priming Low Representativeness condition) and additivity should be observed in the Priming High Representativeness Condition. However, if the availability of an instance is not sufficient, if a narrow interpretation of categories is so ingrained that making atypical instances available in memory does not influence how people conceive of the category being judged, then the priming unpacking conditions should produce additive judgments. That is, we should see no effect of the priming manipulation.

The availability hypothesis predicts a main effect of representativeness and no effect of type of unpacking.

The narrow interpretation hypothesis predicts a Representativeness by Unpacking interaction due to a negative effect of low representativeness in the direct unpacking condition and no other differences.

Method

162 first-year students participated in the experiment, 76 sampled at the University of Durham (UK), and 86 at the University of North Carolina at Greensboro (US). Participants were presented with booklets containing eight examples from one of the 5 experimental conditions, followed by 16 items asking for judgments of representativeness. The judgments of representativeness were obtained as a validation check on the assignment of examples in the high- and low-representativeness conditions.

Results

Representativeness judgments

Table 2 presents mean representativeness estimates for both populations for High Rep and Low Rep conditions. As expected, ratings for "High Representativeness" items were much higher than those for "Low Representativeness" items.

Table 2: Mean Population by Representativeness subjective representativeness estimates.

	High Rep	Low Rep
Greensboro	65.7	34.3
Durham	61.4	30.0

To make sure our UK and US population samples were comparable, we conducted a 2 (Population) by 2 (Representativeness) repeated-measures ANOVA across items. The main effect of representativeness was highly significant ($F(1,14)=23.33$, $p<.001$). There was also a main effect of population, US probability judgments were about 4 percentage points higher than UK judgments ($F(1,14)=4.55$, $p<.06$). Most importantly, no interaction was observed ($F<1$). The results justify the assignment of items to High- or Low-representativeness conditions.

An examination of the judgments for each item showed that the direction of representativeness judgments for one were opposite to our expectations. This item was excluded from subsequent analyses.

Probability Judgments

Population To test whether population influenced probability judgments, we performed a 2 Population by 5 Experimental condition ANOVA across participants. Only Experimental condition reached significance ($F(4,151)=5.18$, $p<.001$). The data for the two populations were combined for subsequent analyses.

Unpacking by Representativeness Table 3 presents mean subjective probability judgments for each Unpacking (direct vs. priming) and Representativeness (high vs. low) condition. The mean of the direct low-representativeness cell is the lowest; means of the other cells are about equal.

Table 3: Mean probability judgments by Type of unpacking and Representativeness.

	High Rep	Low Rep
Direct	57.3	44.6
Priming	55.1	55.0
Packed	56.2	

The data were analyzed by two 2 Unpacking by 2 Representativeness analyses of variance, one by participants (F_1) and one by items (F_2). Unpacking had a significant main effect by participants but not by items ($F_1(1,124)=3.96$, $p<.05$; $F_2(1,6)=1.87$, $p>.22$). Representativeness had a significant main effect by participants ($F_1(1,124)=9.71$, $p<.005$) but only a marginal effect by items ($F_2(1,6)=3.03$, $p<.14$). The interaction was significant by participants ($F_1(1,124)=9.37$, $p<.005$) but only marginally by items ($F_2(1,6)=3.08$, $p<.14$).

Two-tailed t-tests compared each Unpacking by Representativeness condition to the rest. The only tests reaching significance were those comparing the direct low-representativeness condition to each of the others.

Superadditivity Mean probability ratings for each of the four Unpacking by Representativeness conditions were compared to the mean rating for the packed condition (M=56.2) to detect deviations from additivity. The only condition that deviated substantially from additivity was the direct low-representativeness condition that demonstrated superadditivity: $t(61) = 3.27$, $p < .005$ (participants); $t(6) = 2.51$, $p < .05$ (items). The item analysis for the direct high-representativeness condition suggested a small amount of subaddivity: $t(6) = 2.43$, $p < .06$; but $t < 1$ (participants).

Discussion

The present study investigated the hypothesis that people interpret category descriptions narrowly, in terms of representative instances, when making subjective probability judgments by crossing representativeness with type of unpacking. We found a representativeness by unpacking interaction due to a negative effect of low representativeness in the direct unpacked condition. Only the direct low-representativeness ratings substantially deviated from

additivity: they were superadditive. Predictions were confirmed with both British and US samples.

The present data replicated Hadjichristidis et al.'s (1999) finding that directly unpacking unrepresentative instances induces implicit superadditivity. One account of these findings is that unpacked instances are treated as a pragmatic cue for determining what the experimenter means by the category label. When asked about "tarantulas or some other type of pet," people might infer that the experimenter has a different category in mind than when asked only about "pets." This account is indeed consistent with our data, but only if construed in a way equivalent to our hypothesis. The data suggest that people interpret categories narrowly and the explicit inclusion of atypical instances broadens the normal interpretation. But our methodology rules out the interpretation that we are merely asking people to a judge different category in the Low Representativeness condition. In every case, categories were described in the current study by clearly stating the alternative hypothesis (e.g., in Table 1, the judged event is always stated along with "as opposed to loving all pets"). Therefore, although we believe our effect depends on how categories are interpreted, it does not represent a mere task demand induced by pragmatic biases. Rather, it represents a central and generalizable aspect of probability judgment of categorical events.

Support theory Our finding that direct unpacking of low representative instances induces superadditivity disconfirms support theory's prediction that unpacking cannot have a negative effect on probability judgments. Our conclusion is independently supported by Macchi, Osherson, and Krantz (1999) who found that unpacking low-support instances resulted in explicit superadditivity for binary partitions.

Implicit subadditivity is not a robust phenomenon. Rottenstreich and Tversky (1997) themselves predicted it three times, but only observed it twice. Implicit subadditivity obtained in the Trial problem, which pitted the hypothesis "the trial will not result in a guilty verdict" against the disjunction "the trial will result in a not guilty verdict or a hung jury". It also obtained in the Homicide problem, which pitted "death from homicide" against "death from homicide by an acquaintance or a stranger". In both cases, they explained subadditivity in terms of enhanced availability. In the first problem, participants might not have considered the hung jury possibility in the packed condition. In the second, the unpacked hypotheses may have brought a host of possible causes of death to mind (e.g., crimes of passion) that would not have been available in the packed condition. In sum, support theorists have themselves identified a key factor that limits the generality of implicit subadditivity.

We believe that minor modifications would allow Support Theory to capture superadditive probability judgments. Here are some possible changes:

1. Allow for negative support.
2. Stick to nonnegative support, but modify the global support function (make it average rather than summed support).
3. Allow that unpacked instances replace instances that would otherwise have been available at the time of judgment.

Dynamic global support functions? A further possibility is that implicit subadditivity and superadditivity reflect different functional relations between the support attached to packed categories and the support attached to their unpacked instances (global support functions). Subadditivity may involve summing of support across instances, whereas superadditivity may involve averaging of support. In Rottenstreich and Tversky's (1997) examples, support is based on subjective impressions of frequencies or reasons, whereas in our "natural fuzzy category" examples (e.g. pets, restaurants), support is based on similarity. The support for the unpacked hypothesis of the Trial problem, for instance, seems to involve estimating the relative frequency of a "not guilty verdict" and a "hung jury" and adding them up. In contrast, the support for the unpacked hypotheses in our examples seems to involve estimating the similarity of the category instances to the description and averaging them out. Corroborating evidence that the similarity-based global support function may average support comes from Rottenstreich, Brenner, and Sood (1999) who showed that similarity-based likelihood judgment gives rise to nonmonotonicities: the support of a disjunction is less than that of one of its components. They presented participants with a description of Linda: an outspoken, socially conscious, and single woman. "Linda is a journalist" enjoyed higher support than the disjunctive hypothesis "Linda is a journalist or a realtor". Nonmonotonicities cannot be explained by a global support function that adds support, but could by one that averages support.

In sum, the global support function may change dynamically depending on the particular base of support –e.g. objective frequencies, similarity, and reasons. Similarity-based likelihood judgment may involve averaging; frequency-based likelihood judgment may involve summing. Supporting this hypothesis, Rottenstreich et al. (1999) showed that case judgments (e.g., the Linda example) that presumably involve similarity-based reasoning give rise to nonmonotonicities, whereas class judgments (e.g., the probability that a randomly selected American is a

journalist) that presumably promote frequency-based reasoning, do not. These suggestions all stay close to the spirit of Support Theory because we find its assumption that probability judgments are mediated by judgments of evidence to be appealing and worth maintaining in the next generation of descriptive theory (see also Fischhoff, Slovic, & Lichtenstein, 1979).

Decision-making Much everyday decision-making depends on subjective assessments of probability. For instance, the premium one is willing to pay for health insurance depends on a subjective assessment of the likelihood of getting hospitalized for the cases that the health insurance covers (see Johnson et al., 1993). The events for which an insurance provides coverage can be described in many ways. The results we report here can presumably be extended to the domain of decision-making.

Acknowledgments

We want to thank David Over and Rosemary Stevenson for helpful discussions on Support Theory and Marianne Harrison for running pilot studies on this project. Constantinos Hadjichristidis was supported by the ESRC grant No. R000239074 on *Belief Revision and Uncertain Reasoning.*

References

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance, 4,* 330-334.

Hadjichristidis, C., Stibel, J. M., Sloman, S. A., Over, D. E., & Stevenson, R. J. (1999). Opening Pandora's box: selective unpacking and superadditivity. *Proceedings of the European Conference on Cognitive Science* (pp. 185-190). Siena, Italy.

Johnson, E. J., Hershey, J., Meszaros, J., & Kunreuther, H. (1993). Framing, probability distortions, and insurance decisions. *Journal of Risk and Uncertainty, 7,* 35-51.

Kahneman, D. & Miller, D. T. (1986). Norm Theory: comparing reality to its alternatives. *Psychological Review, 93,* 136-153.

Macchi, L., Osherson, D., & Krantz, D. H. (1999). A note on superadditive probability judgment. *Psychological Review, 106,* 210-214.

Rottenstreich, Y., Brenner, L., & Sood, S. (1999). Similarity between hypotheses and evidence. *Cognitive Psychology, 38,* 110-128.

Rottenstreich, Y. & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review, 104,* 406-415.

Tversky, A. & Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business, 59*, 251-278.

Tversky, A. & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101,* 547-567.

On the Evaluation of *If p then q* Conditionals

Constantinos Hadjichristidis (constantinos.hadjichristidis@durham.ac.uk)
Department of Psychology, University of Durham
Durham, DH1 3LE, UK

Rosemary J. Stevenson (rosemary.stevenson@durham.ac.uk)
Department of Psychology, University of Durham
Durham, DH1 3LE, UK

David E. Over (david.over@sunderland.ac.uk)
School of Social Sciences, University of Sunderland
Sunderland, SR1 3SD, UK

Steven A. Sloman (steven_sloman@brown.edu)
Department of Cognitive & Linguistic Sciences, Brown University
Box 1978, Providence, RI 02912, USA

Jonathan St. B. T. Evans (j.evans@plymouth.ac.uk)
Centre for Thinking and Language, Department of Psychology
University of Plymouth, Plymouth PL4 8AA, UK.

Aidan Feeney (aidan.feeney@durham.ac.uk)
Department of Psychology, University of Durham
Durham, DH1 3LE, UK

Abstract

We propose that when evaluating conditionals, people construct an imaginary world that contains the antecedent, and then evaluate the plausibility of the consequent being true in the same world. Thus, when asked for an estimate of the probability of the conditional, people should produce the conditional probability of its consequent given its antecedent. We contrast this view with a view based on the theory of mental models, in which the judged probability of a conditional is derived from the proportion of models in which the premises are true. Study 1 examined this hypothesis by comparing probability estimates for (i) category-based conditional arguments (e.g. *If robins have ulnar arteries then sparrows have ulnar arteries*), (ii) corresponding conditional probabilities in the form of suppositions (e.g. *Suppose you knew that robins have ulnar arteries. How likely would you think it was that sparrows have ulnar arteries?*) and (iii) the argument strength of corresponding inductive arguments (e.g. *Fact: Robins have ulnar arteries. Therefore: Sparrows have ulnar arteries. How convincing do you find this argument?*) All three estimates were highly correlated, a finding that supports our hypothesis. The similarity between the two categories (e.g. robins and sparrows) was also manipulated. Similarity affected all three estimates equally, similar items being given higher estimates than dissimilar items. This finding indicates that similarity is one basis for the plausibility judgements. Study 2 tested our hypothesis using conditional statements with known probabilities. The results favoured our hypothesis. We discuss these results in terms of philosophical and psychological views of conditionals, and suggest that they bring together kinds of reasoning that are traditionally studied separately, such as conditional reasoning, induction, and judgements of probability.

Introduction

Psychological research on inductive and deductive reasoning has traditionally examined reasoning based on premises classified as true. Such research ignores most everyday reasoning, which is based on uncertain premises. Premise uncertainty, in turn, rightly influences the degree of certainty in the conclusion of an inference (e.g. Stevenson & Over, 1995). Understanding everyday reasoning, therefore, involves understanding subjective premise uncertainty, and the way in which such uncertainty gets translated into uncertainty about the conclusion of an inference. The present article investigates subjective uncertainty about conditional premises of the form *If p then q*.

The article focuses on the way in which people evaluate conditional arguments and how they arrive at judgements of the probability of a conditional. We propose that people evaluate conditionals with reference to imaginary situations that they mentally

construct; in particular, people evaluate the plausibility of the consequent in an imaginary situation in which the antecedent is true. For example, on encountering the conditional *If you study hard then you will pass the exam,* we propose that reasoners mentally construct an imaginary situation in which they study hard and then pass the exam. This imagined situation may be judged more plausible than one in which they study hard and do not pass the exam. Such a judgement might be mediated by causal schemas, e.g. one's intuitive theories about studying and success or failure (Collins & Michalski, 1989). In other situations, similarity (Osherson et al, 1991) or a judgmental heuristic, such as representativeness or availability (Kahneman et al, 1982) might be used. If people do indeed evaluate conditionals in the above manner, then when asked to estimate the probability of the conditional, they should state the probability of the consequent given that the antecedent is true; that is, they should give the conditional probability.

The view presented above is similar to Ramsey's (1931) notion of how a conditional is evaluated, and also has some similarities to proposals made by Adams (1975) and Edgington (1995). Ramsey's idea was that when we evaluate a conditional, we add the antecedent to our stock of beliefs, leaving everything else as undisturbed as possible, and then examine whether our new stock of beliefs contains the consequent. Our proposal that people construct an imaginary world that contains the antecedent is comparable to updating one's knowledge base by adding the antecedent and making the minimal changes resulting from the presence of the antecedent. People then assess the likelihood that the consequent also holds, using either heuristics or sometimes beliefs about relative frequencies. The psychological validity of this "imaginary worlds" view of conditionals has not yet been tested, although conditionals have been linked to conditional probabilities in other psychological work (Stevenson & Over, 1995; Oaksford, Chater, & Larkin, 2000). The present experiments test this view.

The mental models theory provides a contrasting view of how individuals untrained in logic evaluate the probability of conditional statements. Johnson-Laird et al. (1999) propose that such individuals infer the probability of events by reasoning extensionally. They construct mental models representing true possibilities (the *principle of truth*) and estimate the sum of the probabilities of the models in which the event occurs. *If p then q* conditionals are understood by representing up to the following three explicit mental models (¬ stands for the negation of a premise):

p q
¬ p q
¬ p ¬ q

"p ¬q" is not represented because it is a false possibility, though it can be inferred as the complement of the fully explicit models, although this rarely happens (see Barres & Johnson-Laird, 1997). However, consideration of the false possibility is critical for the conditional probability interpretation of conditional statements; the conditional probability of q/p depends on the relative ratio of *pq* to *p¬q* possibilities, i.e. Pr(q/p) = Pr(pq)/ [Pr(pq)+Pr(p¬q)]. Therefore, evidence for a conditional probability interpretation of conditional statements would challenge the mental models theory.

Study 1: Subjective probabilities

In Study 1 we compare the imaginary world hypothesis with the mental models hypothesis by obtaining judgements of (1) the probabilities of conditional arguments, (2) conditional probabilities, and (3) judgments of the convincingness of inductive arguments, that is, judgements of argument strength. Examples of the materials are shown in Table 1.

We propose that when asked to estimate the probability of the conditional shown on the top panel of Table 1, participants will evaluate the conditional in the same way as they evaluate the conclusion of the conditional probability statement shown in the middle panel of Table 1. That is, the reasoning in both cases will be based on the same representation, an imaginary world in which horses have stenozoidal cells and in which the plausibility of cows having stenozoidal cells is assessed.

We also propose that participants evaluate inductive arguments, like the one in the last panel of Table 1, in a similar way. Inductive argument tasks ask participants to assume that p is a fact. Our hypothesis, that when judging the probability of conditional, people imagine a world in which p is true and make judgements about that world, predicts that they should give the same judgement as they give when explicitly told that p is in fact true (i.e., when making argument strength judgements).

Because an imaginary world in which both horses and cows share a property is more representative of the real situation than a world in which horses have the property but cows don't, we expect all three types of judgments to be relatively high. Furthermore, we expect the probability of the conditional to be highly correlated with the conditional probability judgements on the one hand and judgements of argument strength on the other, since we argue that they all measure the same process. Note that an association between argument strength and conditional probability judgements has been presupposed in psychological research (e.g. by Sloman, 1998). By contrast, the mental models view of conditionals does not consider the case in which horses

have the relevant property but cows do not. Consequently it would not predict that judgements of the probability of the conditional would be highly related to either conditional probability judgements or argument strength judgements.

Table 1. Study 1: An example of materials used in Study 1. [Note: The example is from the similar condition. Half of the materials were in the dissimilar condition.]

Probability of conditional condition
Peter said the following: If horses have stenozoidal cells, then cows will have stenozoidal cells. How likely do you think it is that what Peter said is true?
0 1 2 3 4 5 6 7 8 9 10
not at all likely very likely
Conditional probability condition
Suppose you knew that horses have stenozoidal cells. How likely would you think it was that cows have stenozoidal cells?
0 1 2 3 4 5 6 7 8 9 10
not at all likely very likely
Inductive argument condition
Fact: Horses have stenozoidal cells

Conclusion: Cows have stenozoidal cells
0 1 2 3 4 5 6 7 8 9 10
not at all convincing very convincing

Study 1 also examined the hypothesis that the similarity between the two categories mediates the judgments in all three conditions. The more similar the categories, the more structure (features and dependencies) their representations share. Thus people are likely to infer that the more known structure two categories share, the more novel structure they are likely to share. We expect, for instance, an imaginary situation, in which similar categories (e.g. cows and horses) share a novel property, will be judged more plausible than an imaginary situation, in which dissimilar categories (e.g. cows and mice) share a novel property. (See Osherson et al, 1991, for a model of how conditional probabilities can be derived from similarity judgements.) Consistent with our view, research on category-based inductive arguments (like the one in the last panel of Table 1) has shown a robust effect of similarity (see e.g. Rips, 1975; Sloman, 1993). Moreover, to the extent that such similarity-based reasoning is non-extensional (see Johnson-Laird et al., 1999), it falls outside the scope of mental models theory, which only considers extensional reasoning.

Method

Participants. Forty-one first-year psychology students volunteered to participate in this study.

Design. Type of Measure (probability of conditional vs. conditional probability vs. argument strength) was crossed with Similarity (similar vs. dissimilar category pairs) in a mixed design with repeated measures on the last factor.

Procedure. Participants were presented with booklets containing 18 examples in one Type of Measure condition. Half of the examples in each condition contained similar and half dissimilar mammal pairs. The assignment of category pairs to similarity conditions was controlled by an independent group of twelve participants who were asked to rate the biological similarity of the 16 mammal pairs in a 0-10 scale, were 0 was labeled as "highly dissimilar" and 10 as "highly similar." The mean ratings for the similar and dissimilar items were, respectively, 7.39 (min=5.92, max=8.92) and 1.74 (min=.92, max=2.33). The results therefore justify the assignment of items to the similar or dissimilar conditions.

Participants in the probability of the conditional condition (N=16) were told that they would be presented with statements uttered by a person. Their task was to say how likely they thought it was that what the person said was true on a 0-10 scale, where 0 was labeled as "not at all likely" and 10 as "very likely." We used this task to obtain judgements of the probability of the conditional to ensure that our instructions did not encourage participants to give conditional probability judgements for superficial reasons[1]. If participants were simply asked "How likely do you think it is that *If p then q?*" they might interpret the question as asking the question "If p, what is the probability that q?" That is, as a direct request for the conditional probability of the consequent given the antecedent. This problem arises because a conditional consists of a main (the consequent) and a subordinate (the antecedent) clause, and it has been shown that, when processing sentences containing main and subordinate clauses, people often assume that the subordinate clause is true (Baker & Wagner, 1987). Our instructions were designed, therefore, to avoid responses based on this kind of linguistic paraphrase and to ensure instead that they were based on a conceptual understanding of the conditional.

Participants in the conditional probability condition (N=10) were told that they would be presented with examples asking them to suppose that a statement is

[1] We thank Phil Johnson-Laird and Vittorio Girotto for suggesting these instructions for the framing of the conditional probability condition.

true. Based upon this supposition, they had to judge the likelihood that a second statement is true. The same scale was used as for the Probability of the conditional participants. Participants in the argument strength condition (N=15) were told that they would be presented with a series of arguments, each containing a fact (which should be taken as true) separated from a conclusion by a line. Their task was to describe how convincing they found each argument on a 0-10 scale, where 0 was labeled as "not at all convincing" and 10 as "very convincing." Participants in all conditions worked through examples similar to the test items before starting the experiment.

Results and Discussion

Correlation statistics Table 2 presents the mean correlation coefficients relating the three types of measures across items. As predicted by the imaginary worlds view, the three measures were significantly correlated (beyond the .001 level).

Table 2. Mean correlation coefficients by items for each of the three conditions. CP=conditional probability. PC=probability of conditional. AS=argument strength.

	PC	*CP*	*AS*
PC	*1.0*	*.99*	*.94*
CP		*1.0*	*.96*
AS			*1.0*

Similarity Table 3 presents mean Type of measure by Similarity estimates. In each Type of measure condition, ratings for similar items were higher than ratings for dissimilar items.

Table 3. Mean Type of Measure by Similarity estimates CP=conditional probability. PC=probability of conditional. AS=argument strength.

	Similar Items	Dissimilar Items
PC	5.78	2.76
CP	6.66	2.83
AS	5.40	2.30

The data from each measure were analyzed by pairwise t-tests for participants, and independent t-tests for items. Pairwise tests were used in preference to a single ANOVA because we cannot assume that the three measures are comparable. For each type of measure, both across participants and items, similarity had a significant effect (beyond the .005 level). These results suggest that the plausibility judgements underlying the imaginary worlds view can be influenced by similarity. The mental models view, however, cannot account for either the correlational results or the effect of similarity.

Study 2: Objective probabilities

Study 2 also investigated how people evaluate conditional statements but with conditionals of known conditional probabilities. The use of known probabilities provides a direct test of our two competing hypotheses, because it allows judgements about the probability of the conditional to be directly compared with the objective conditional probability.

Participants were given three different versions of a text describing a probability problem. For example, a third of the participants read the following text and were then asked to estimate the probability that what Peter said was true:

> In an effort to boost its image, Waterstones bookstore organised lotteries in several Primary schools in Durham. In each school, only the 10 best students participated in the lottery. The name of each participant was written on a piece of paper and was put in a hat. A blindfolded teacher drew a piece of paper from the hat. The student whose name was written on that paper won an autographed storybook. In Durham Gilesgate Primary School the participants were 8 boys and 2 girls. A piece of paper was drawn from the hat. Peter, the father of one of the participants, cannot see the winner's name but says: "If a boy has won the lottery, then my son won it."

According to the imaginary world hypothesis, participants should construct an imaginary situation in which a boy wins the lottery and then consider how likely it is that the boy is Peter's son. Since there are 8 boys all together, this conditional probability is 1/8.

According to the theory of mental models, the correct answer depends upon considering the fully explicit models of the proposition and finding the proportion of models in which the proposition is true. These explicit models, which represent the true possibilities, are shown below, with tags indicating their relative frequencies. (*Boy* stands for the antecedent; *Son* stands for the consequent).

Boy	Son	1/10
¬Boy	Son	0[2]
¬Boy	¬Son	2/10

The proportion of models, therefore, in which the proposition is true is 3/10. We call this the material implication (MI) evaluation of the conditional.

[2] No doubt participants will rule out the possibility of ¬*Boy and Son* on pragmatic grounds. However, for the above problem, this does not affect the predicted probability judgement.

The probability estimates that agreed with one or other of these two evaluations (the conditional probability or material implication) were coded as supportive of either the imaginary world view or the mental models view respectively.

Method

Participants Forty-eight first-year undergraduate volunteers participated in Study 2. The sample included the same forty-one students that participated in Study 1.

Procedure Each participant was presented with a booklet containing one version of the problem given above. In one version the sample of children consisted of 2 boys and 8 girls (the 2b-8g version), in a second of 5 boys and 5 girls (the 5b-5g version), and in a third of 8 boys and 2 girls (the 8b-2g version). Table 4 lists the predictions for the imaginary world and the mental models view for each of the three versions of the problem.

Table 4. Conditional probability and material implication predictions for each of the 3 versions of the problem.

	Imaginary world view (Conditional probability)	Mental models view (Material implication)
2b-8g version	1/2	9/10
5b-5g version	1/5	6/10
8b-2g version	1/8	3/10

Results

Table 5 presents the number of participants in each version of the problem whose response agrees with one of the two evaluation modes for each version. Out of the 48 participants, 44 gave numerical answers.

Table 5. Number of participants in each version whose response falls in one of the two evaluation modes. N=number of participants who gave a numerical response for each version.

	N	Imaginary world View	Mental models view
2b-8g version	13	9	0
5b-5g version	14	5	0
8b-2g version	17	9	1
Total	*44*	*23*	*1*

Out of those 44, 24 gave a response that could be classified in one of the two response modes. The results reported in Table 5, therefore, account for 55% of those responses. The results of all except one of these 24 participants agreed with the conditional probability evaluation (X^2 = 20.17). These data clearly favor the imaginary worlds view of conditionals over the mental models view.

The main numeric responses made by the remaining 20 participants were "1/10" or its arithmetic equivalent (N=6), "1/2" or its arithmetic equivalent (N=7), "2/10" or arithmetic equivalent (N=5)[3]. The "1/10" responses are consistent with the mental models view that participants represent an explicit model of the premise and ignore other possibilities. The remaining two responses may reflect failures to understand the conditional.

General Discussion

These results with both subjective and objective probabilities support the imaginary worlds view. Judgements of the probability of a conditional correlated highly with conditional probability judgements and argument strength judgements in Study 1, and they matched the objective conditional probabilities in Study 2. Mental models theory cannot explain these results because it only considers true possibilities. Even if we grant that mental models theory allows that the false possibility may be inferred, the theory still cannot explain our results because it has no mechanism for calculating conditional probabilities when presented with a conditional (see Johnson-Laird et al, 1999).

Furthermore, mental models theory fails to explain the similarity effect found in Study 1. By contrast, this effect follows from the imaginary worlds view, which claims that reasoners evaluate conditionals by representing the antecedent and consequent in an imaginary world and then evaluating the plausibility of this world. Since similarity is a key component of plausible reasoning (Osherson et al, 1991), it follows that similarity should also be a key component in judgements of the plausibility of the consequent being true in a world in which the antecedent is true.

The results of Study 2 also suggest that explicitly presenting the negated antecedent is, in itself, insufficient to promote its inclusion in a mental representation. People might represent such possibilities when background knowledge makes them salient. For example, "If John is in Paris then he is in France" might make people represent the possibility "If John is not in Paris then he is not in France." But as far as the basic evaluation of a conditional is concerned, our results

[3] Some of the responses in the 2b-8g version that are classified as "conditional probability" responses may in fact be "fifty/fifty" responses. However, even if the 2b-8g version is omitted from the analysis, the results still clearly favor the conditional probability view.

suggest that people construct an imaginary world in which the antecedent holds and then consider the likelihood that the consequent holds in the same world.

Our notion of imaginary worlds could be seen as an example of mental models. However, our results suggest that mental models are represented and deployed in ways other than those proposed by Johnson-Laird (e.g. Johnson-Laird et al, 1999) when evaluating the probability of conditionals. For example, the principle of truth cannot apply to uncertain conditionals, since the "false possibility" (p ¬q), must be at least implicitly considered to arrive at the conditional probability. Furthermore, the role of similarity in evaluating an uncertain conditional needs to be included in such a theory.

Our proposal has something in common with possible worlds analyses of ordinary conditionals (Stalnaker, 1968; Lewis, 1973). However, if these conditionals are analyzed in this way in formal semantics, then there are technical reasons why the probability of a conditional cannot be absolutely identified with the corresponding conditional probability. (See Jackson, 1991, for the main technical papers on this issue.) But the technical issue not withstanding, there is reason to hold that the assertion and evaluation of most ordinary conditionals will make them closely related to the corresponding conditional probabilities (Stevenson & Over, 1995; Edgington, 1995). This is all we need for our psychological claims here. Our view is that the judged probability of an ordinary conditional will usually be estimated by assessing the plausibility of the consequent being present in a model that contains the antecedent. Finally, since our views derive from philosophical accounts of conditionals, the present studies also provide a bridge between philosophical and psychological accounts of If p the q conditionals. They also bring together components of reasoning that have been traditionally studied separately such as conditional reasoning, induction, and judgements of probability.

Acknowledgements

The present research was supported by ESRC grant No. R000239074 on *Belief Revision and Uncertain Reasoning*. We thank Vittorio Girotto and Phil Johnson-Laird for helpful discussions and an anonymous reviewer for very useful comments.

References

Adams, E.W. (1975). *The logic of conditionals.* Dordrecht: Reidel.

Baker, L. & Wagner, J.L. (1987). Evaluating information for truthfulness: The effects of logical subordination. *Memory & Cognition, 15,* 279-284.

Barres, P.E.& Johnson-Laird, P.N. (1997). Why is it hard to imagine what is false? *Proceedings of the 19th Annual conference of the Cognitive Science Society,* pp. 475-478.

Braine, M.D.S. & O'Brien, D.P. (1991). A theory of If: A lexical entry, reasoning program, and pragmatic principles. *Psychological Review, 98,* 182-203.

Collins, A. & Michalski, R. (1989). The logic of plausible reasoning: a core theory. *Cognitive Science, 13*, 1-49

Edgington, D. (1995). On conditionals. *Mind, 104*, 235 329.

Grice, H.P. (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.), *Syntax and semantics. Vol. 3 Speech acts.* New York: Seminar Press.

Jackson, F. (Ed.) (1991). *Conditionals.* Oxford: Oxford University Press.

Johnson-Laird, P.N. (1983). *Mental Models*. Harvard University Press, Cambridge, Mass.

Johnson-Laird, P.N. & Byrne, R.M.E. (1991). *Deduction.* Hove: Erlbaum.

Johnson-Laird, P.N. & Byrne, R.M.E. (2000). Conditionals: A theory of meaning, pragmatics and inference. Unpublished manuscript.

Johnson-Laird, P.N., Legrenzi, P., Girotto, V., Legrenzi, M., & Caverni, J-P. (1999). Naive probability: a mental model theory of extensional reasoning. *Psychological Review, 106,* 62-88.

Kahneman, D., Slovic, P. & Tversky, A. (1982). *Judgements under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Lewis, D.K. (1973). *Counterfactuals*. Basil Blackwell.

Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory & Cognition, 26,* 883-899.

Osherson, D.N., Stern, J., Wilkie, O., Stob, M, & Smith, E.E. (1991). Default probability. *Cognitive Science, 15,* 251-269.

Ramsey, F.P. (1931). *Foundations of Mathematics*. Routledge Keagan Paul.

Rips, L. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning, and verbal Behavior, 14,* 665-681.

Sloman, S.A. (1993). Feature-based induction. *Cognitive Psychology, 25,* 231-280.

Sloman, S.A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology, 35,* 1-33.

Stalnaker, R.C. (1968). A Theory of Conditionals. In N. Rescher (Ed.) *Studies in Logical Theory, American Philosophical Quarterly* monograph series. No. 2, 98-112. Basil Blackwell.

Stevenson, R.J., & Over, D.E. (1995). Deduction from uncertain premises. *The Quarterly Journal of Experimental Psychology, 48A (3),* 613-643.

Very Rapid Induction of *General* Patterns

Robert F. Hadley
School of Computing Science
and Cognitive Science Program
Simon Fraser University
Burnaby, B.C., V5A 1S6
hadley@cs.sfu.ca

Abstract

Marcus (1998) and Phillips (2000) each have produced examples of human *generalizations* which, they argue, cannot be matched by the best known connectionist architectures and training algorithms. However, I argue that humans perform the crucial generalizations *without being trained* on the exemplars that Marcus and Phillips cite. So, in a sense, the issue whether networks can be trained to perform the crucial generalizations is a red herring. I argue further that humans achieve the dramatic generalizations in question as a side-effect of a variety of *pre-existing* skills, working in concert. Finally, it is shown that the "hard cases" displayed by Marcus and Phillips do in fact provide the basis for a serious challenge to *pure* (non-modular) connectionist architectures.

Introduction

In Marcus (1998, in press) and Phillips (2000), intriguing refinements on the (1988) Fodor-Pylyshyn "generalization challenge" are presented. Both Marcus and Phillips argue that linguistically competent humans exhibit important forms of *generalization* that backpropagation-trained networks (both recurrent and feedforward) cannot attain. Though neither author categorically asserts that their negative conclusions apply to every form of connectionist training, both authors argue that commonly recognized varieties of *eliminativist* architectures (i.e., those eschewing classical representations) are at stake.[1]

In this paper, I examine two instances which typify the "hardest challenges" produced by these authors. While I agree that they have each exposed some important training limitations of backpropagation networks, I shall argue that humans perform the crucial generalizations *without being trained* on the exemplars that Marcus and Phillips cite. So, in one sense, the issue whether networks can be *trained* to perform the crucial generalizations is a red herring. As I argue, humans possess the relevant generalization capacity because they have previously acquired separate skills which, working in concert, allow for nearly instantaneous pattern induction and reasoning. To be sure, the prior acquisition of these separate skills may involve "training up" various sub-networks in our brains, but this prior training may well *not* involve the "hard" kinds of generalization at issue.

Having said that, I would emphasize that my eventual, general conclusion supports both the positions of Marcus and Phillips. For, in the final section I consider whether connectionist architectures are capable (without implementing classically symbolic methods) of orchestrating the application of our "prior skills" in a fashion that permits very rapid pattern induction and reasoning. My conclusion favors the classicist position on this issue. Moreover, I propose a new challenge for eliminative connectionists which, in my view, formulates the deeper difficulty posed by the aforementioned "hard cases" of Marcus and Phillips.

The Hard Generalization Tasks

Generalizing Outside the Training Space

Marcus (1998) defines a network's training space as the N-dimensional vector space created by the non-zero training values of the N units comprising the network's input array. A datum presented during the network's (post-training) test phase lies "outside the training space" if and only if that datum does not fall within the vector space just mentioned. In effect, this entails that the datum is *novel* relative to the training corpus. For example, any datum would be novel in the relevant sense if it presented non-zero values to the input array in units that contained only zero values during training.

This "generalization hurdle" differs somewhat from the hierarchy of systematicity given in Hadley (1994a), but it appears equivalent to one of several levels of generalization formulated in Niklasson and van Gelder (1994). Interestingly, in the latter paper, the authors claim to satisfy this particular generalization challenge. Their claim is questioned in Hadley (1994b) and wholly disputed by Marcus (1998). Moreover, Marcus discusses a number of specific ways in which a network can fail to generalize outside its training space, and we now consider a particular "hard case" which Niklasson and van Gelder had not addressed.

Suppose a linguistically competent human is presented with the following series: "A rose is a rose", "A frog is a frog", "A pencil is a pencil". Humans will typically have no difficulty inducing the general pattern and complet-

[1] Here I am following Fodor and Pylyshyn's (1988) usage, according to which a composite (or complex) representation is classical in structure if one cannot activate (or token) that representation without, at the same time, tokening its syntactic constituents.

ing the following sentence: "A blicket is a ...". Humans will succeed here even though 'blicket' is a novel word which is outside their training space. In contrast, Marcus offers persuasive arguments, based upon the training-independence of output nodes, to show that backpropagation networks necessarily fail to match this success. These arguments are buttressed by several connectionist experiments conducted by Marcus.

On the basis of the above and related tasks, where a strong discrepancy exists between human performance and that of eliminative networks, Marcus concludes **C**: that human success in such cases is *not purely* due to any training of putative eliminative networks within our brains. This conclusion forms a keystone of Marcus' larger thesis – that the human ability to discover general patterns in cases such as these involves symbolic rule induction, and the application of such rules entails variable binding.

Now, while I agree with conclusion (**C**), I accept this conclusion for reasons other than any offered by Marcus. For one thing, I suspect that some Hebbian-competitive networks *can* generalize outside their training spaces on at least some tasks. This suspicion derives from recent experimentation with an architecture I have reported in (Hadley, et al, to appear). Another difficulty is that Marcus himself notes that when *distributed*, rather than local, representations are assigned to input tokens, backpropagation networks will, at first blush, provide the appearance of generalizing outside their training spaces. For example, in the "A blicket is a .....' test, a backpropagation network can successfully produce the distributed representation for 'blicket', *provided all the separate features* encoding blicket had, at some point, been employed in various nouns during the training phase. Admittedly, one could argue that this last proviso undermines any well founded claim to generalization outside the training space, but in doing so, one would undercut the entire force of the 'blicket' test case. For the word 'blicket' itself possesses only phonetic and graphemetic features that humans have often encountered prior to being presented with the 'a blicket is a ...' test phrase. That is, a plausible distributed representation for 'blicket' does not contain any features novel to English-speaking humans.

Marcus himself does not stress the objection I have assigned to some anonymous "one". Rather, he primarily objects (Marcus, in press, appendix 1) to the use of distributed representations on the grounds that they fail "... to unambiguously represent all and only the possible continuations to a given string ...". That is, when both nouns and verbs share several features in common (as indeed they would if we employ phonetic or graphemetic features), we run into the *superposition catastrophe* (crosstalk). (I would argue, however, that there is, *at most*, very little overlap between *semantic* features belonging to nouns and those belonging to verbs. For this reason, among others, the system described in Hadley et al, 2000, employs semantic features.) Be that as it may, it remains true that the phonetic and graphemetic features of 'blicket' are not novel. Moreover, *those* features are shared by both nouns and verbs, and, being a nonsense word, 'blicket' has no semantic features. So, if Marcus objects to the deployment of distributed representations in these network experiments, it seems incumbent upon him to demonstrate that humans are using only *local* representations when they successfully generalize from "A rose is a rose", etc. to "A blicket is a blicket". In the absence of such a demonstration, there seems no reason to grant that humans are in fact generalizing outside their training space in cases such as this.

For all the above reasons, I have serious reservations about Marcus' argument for conclusion **C**. Nevertheless, as mentioned, I believe there is a compelling reason to accept (**C**). And, if I am right about this latter reason, then the disputed capacity of eliminative networks to generalize outside their training spaces may be irrelevant *as the task is presently formulated*.

Here is the situation: humans clearly are able to perform very rapid pattern induction, not only in the various cases that Marcus cites, but in many other instances. In the above case, humans are able to induce a general pattern, and supply 'blicket' in response to the test phrase "A blicket is a ... ", within mere seconds after hearing "A rose is a rose", and the few remaining sample sentences. Given the very short time span involved, we may be quite certain that human success in this and similar cases does not stem from some extremely rapid training of "neural networks" (whether eliminative or not). As emphasized in Hadley (1993), in cases where humans make virtually instantaneous inferences, and when they acquire general rules in a matter of mere seconds, rapid synaptic weight change can be ruled out. Synapses simply do not grow fast enough to permit the acquisition of coherent functionality within the span of a few seconds. Functionally coherent synaptic changes occurs within spans of hours or days, not in a few seconds.

Now, it might be objected that in the case of the 'blicket' generalization, humans have in fact had entire days or even years to "train up" their networks, since, arguably, they have frequently heard phrases of the precise form, "an X is an X", in the past. However, this objection falters when we reflect that English-speaking adults have no difficulty inducing a *novel* pattern, and completing the final "sentence" in the following series: "Rose biffle biffle zarple zarple rose"; "Frog biffle biffle zarple zarple frog"; "Blicket biffle biffle — — —". In this case, the pattern being induced is clearly novel, since the pattern (template) itself not only includes the words 'biffle' and 'zarple', but involves a "syntax" that employs a double repetitive pattern not found in English. Yet, humans perform this induction in mere seconds. We must conclude, therefore, that the ability to perform rapid pattern inductions of this kind does not derive from some instantaneous training of a neural network, but must rely on at least some pre-existing skills. Certain of these prior skills involve the capacity to recognize phonemes or graphemes, which doubtless entails modification of synaptic "weights", which in turn (presumably) amounts

to the training of sub-networks within the brain.

Note, however, that this *prior* network training is not specifically directed to the generalization task just considered. The *novelty of the pattern* being induced ensures that very rapid, successful induction of this pattern must arise as a side-effect of prior skill acquisition. An appropriate challenge, therefore, for eliminative connectionism, is not whether a single network can be trained to generalize successfully from the few samples of data cited above, but *whether an essentially non-classical network can exercise its hitherto acquired skills* in a manner that yields, **as a side-effect**, the kind of rapid pattern induction considered above. Clearly, these are deep waters; I shall return to this issue in section 3.

Generalization in Rapid Inference

We turn now to consider an apparently "hard case", presented by Phillips (2000). This case is one of a series of generalization tasks considered by Phillips. Each task in the series possesses features which, at first blush, render it unlearnable by backpropagation methods in feed-forward and recurrent networks. However, Phillips engages in a dialectical process in each case, and *seems* to conclude that, provided overlapping distributed representations are assigned to functionally similar atomic constituents within the input data, then, with one exception, each task becomes learnable. The apparent exception is discussed below.

It is noteworthy, though, that even in the case of this seeming exception, Phillips describes a network capable of performing the task. He produces a carefully designed, fragile (and hand-crafted) network whose prescribed weight configuration suffices to display appropriate generalization behaviour. However, Phillips neither argues that the requisite weights could be acquired by learning, nor that the network possesses any cognitive plausibility. Given the precise and fragile nature of the requisite weight vectors, it seems unlikely that the particular network Phillips discusses could in fact be engendered through training.

Presently, I consider details of Phillips "recalcitrant case", but before doing so it will be helpful to consider a partially analogous example. Let us assume that Fiffle, Giffle, and Kiffle are names of propositions that have truth values. (I assume these three names, *qua* names, are novel for most readers.) Also suppose that the following three statements are true.

> If Fiffle is true, then Giffle is true.
>
> If Giffle is true, then Kiffle is true.
>
> If Kiffle is true, then Fiffle is true.

Finally suppose that Kiffle is true. What else can then be known to be true? Before reading further, I invite the reader to discover what can be inferred.

Doubtless, without effort, you have rapidly inferred the truth of the two remaining propositions, Fiffle and Giffle. Any number of literate humans, who have no training in formal logic, could similarly succeed at this task. Clearly, in the elapsed time between your having read the problem statement and your having derived the remaining propositions *no neural network was trained* within your brain to perform the relevant inferences. Rather, your success stems from a prior ability to engage in iterative processing and *modus ponens* inferences. Arguably, in the case of humans who lack formal logic training, the latter capacity derives from prior training in language use (with sentences of the form: if P then Q).

Of course, from a connectionist perspective, the capacity to apply inference patterns to novel data (Fiffle, Giffle, and Kiffle) is a significant achievement, and it is questionable whether any cognitively plausible ANN experiment has succeeded in this task.[2] However, just as in the case of 'blicket', 'Fiffle', 'Giffle', and 'Kiffle' possess only *non-novel* phonetic and graphemetic features. Given that Marcus was able to train a simple recurrent net to predict 'blicket' in the 'a Y is a Y' formula, there would seem no obstacle, in principle, to the *modus ponens* inference pattern being applied to nonsense words, provided the latter are represented by distributed representations of the right kind.

With this in mind, we now consider the problematic case that Phillips describes, viz., *transverse patterning*. Phillips defines transverse patterning as follows:

> Transverse patterning is an example of a stimulus-response task that depends on *between constituent* relations (my emphasis) . A task instance or problem set consists of three unique patterns (e.g., strings, shapes, etc.) *A, B* and **C**, such that: A predicts B; B predicts C; and C predicts A. Once the transverse patterning task structure has been learnt from the first few problem sets, subjects require only one of the three stimulus-response pairs to predict the remaining two, for any new transverse patterning problem set.

At first glance, there may appear to be an ambiguity in the last of the sentences just quoted. However, carefully read, the sentence tells us that human subjects can predict, when given a single *novel* stimulus-response pair (of the form "shape X predicts the appearance of shape Y") what the two remaining novel S-R pairs, having this general form, will be. In personal communication with Phillips I have verified that the sentence is *not* describing human predictions of the two remaining geometric shapes, given the first geometric shape.

Phillips goes on to relate that *trained* feed-forward and recurrent networks are not able to match the impressive kind of generalization just described. This is not surprising. What is surprising, initially, is that *humans can* predict what the two specific novel S-R pairs will be, given exposure only to one of the three novel S-R pairs. This surprise evaporates, though, when we learn (as I did in further personal communication with Phillips) that human subjects are told in advance what the three geomet-

[2] From a cognitive standpoint, I have serious qualms about Boden's and Niklasson's (2000) recent results on this issue.

ric shapes will be in the novel test situation. Given this, and given that the subjects will have learned the overall structure of the training experiment (following their first few sessions), they are able to *reason* analogically, and to derive by a process of elimination, what the remaining two S-R pairs must be. For example, in the new test situation, subject Kim learns that the novel shapes will be a star, an ellipse, and a hexagon. After being presented with the first S-R pair, Kim is able to *infer* immediately, that (say) A corresponds to the star, and that B corresponds to the ellipse. Knowing this, Kim can reason analogically that the ellipse (corresponding to B) must predict the third geometric shape, the hexagon. Reasoning further, again by analogy, Kim discovers that the hexagon (corresponding to C) must be the predictor of (A), the star.

Now, the crucial point to realize here is that human success in this task involves powerful reasoning skills (both analogical and reasoning by elimination) which the human possessed *prior to any* of the S-R conditioning induced in Phillips experiment. In all likelihood, these prior reasoning skills reside in separate modules which were unaffected by the S-R reasoning presently being considered (see Hadley, 1999, for arguments on the modularity issue). In contrast, the non-modular feedforward and recurrent networks which Phillips contrasts with the human success, possess no prior skills in reasoning of any kind, much less the powerful reasoning capacities that humans bring to the experiment.

The situation is complicated, and confused, by the fact that various of Phillips' remarks create the impression that he is contrasting a human ability to generalize an inference pattern, *which has been acquired in the S-R conditioning sessions*, with an inability, on the part of widely used connectionist architectures, to exhibit comparable generalization. At various points, Phillips explicitly states that the transverse patterning task amounts to the task of *generalizing* logical inference patterns. For example, he says,

> Under controlled conditions, subjects consistently make inferences implied by the underlying logical rules (Halford *et al*. 1998a). Indeed such tasks are ideal tests for systematicity in connectionist networks (Phillips and Halford 1997, Phillips 1999).

Given the type of S-R conditioning employed in Phillips experiment, one naturally supposes that the 'underlying logical rule' that Phillips currently has in mind is tantamount to the rule of *modus ponens* employed in the example I offered above. However, as we have now seen, the crucial human success that Phillips highlights in *not* dependent on a simple application of a given inference pattern (or even repeated applications of that pattern) as occurs in the Fiffle example I provided. Rather it depends upon the composition of prior reasoning skills (a composition involving both analogical reasoning and deduction by a process of elimination) *combined with* an ability to extend inference patterns to novel data.

Phillips believes that his transverse patterning case demonstrates that similarity in distributed representations (of atomic constituents) does not suffice to enable certain kinds of networks to generalize a particular kind of inference patterns to novel data. To the contrary, I have argued that Phillips has conflated the challenge of having a network generalize the application of a single inference pattern with several larger issues. While I certainly agree that the use of distributed representations cannot compensate for the absence of separate, previously acquired reasoning skills (together with the considerable prior training that would engender those skills), this tells us nothing about the efficacy of deploying distributed representations when attempting to apply a *single* known inference pattern to novel data. It is crucial to realize that, in the "transverse patterning" experiment discussed above, humans are doing far more than generalizing the application of a single inference pattern to novel data. They are engaged in an elaborate process involving meta-observations and the composition of separate, sophisticated inference skills.

Moreover, it is questionable whether the S-R training sessions have *trained* human subjects in any *new* inference pattern at all. It seems more likely that the sessions merely provided opportunities for subjects to acquire the base atomic facts (of the form X predicts Y, analogous to the simple "if-then" premises in my *modus ponens* example) which provide fodder for the capacity of humans to apply pre-existing inference skills to novel data.

In any case, I believe it is clear that Phillips' "hard case", like that of Marcus, involves the composition and application of pre-existing skills.

Discussion

In the foregoing, I have argued that, for the generalization tasks in question, the challenges posed to eliminative connectionism have not been felicitously formulated. For, in the tasks considered, we have seen that human success gives every appearance of either arising through the composition of multiple prior skills (viz., language comprehension, analogical reasoning, and deduction by process of elimination, in the case of transverse patterning) or arising as a side-effect of the capacity for language processing (as in the case of 'a blicket is a ...'). Human success in these cases is clearly *not* due to some virtually instantaneous "training" of our synaptic weights. I submit, therefore, that the fundamental challenge posed by these "hard cases" should be formulated essentially along the following lines:

> Demonstrate that a *single* holistic ANN, deploying eliminativist, non-classical representations could, as a manifest *side-effect* of its prior training, perform successfully on either of the "hard" tasks we have considered here.[3]

[3] I regard a network's success on a task, T, as a manifest side-effect of prior training just in case the following two conditions hold: (1) it is clear that prior training had in some way contributed to the success; (2) the network's prior training in-

It might now be objected that the challenge just formulated is unfair, because my wording clearly precludes any solution founded upon the *interactions* of multiple connectionist modules. However, solutions predicated upon the interactions of separate connectionist modules represent a radical departure from the pure connectionist paradigm. Such modular architectures share much in common with traditional, symbolic AI approaches to induction and problem solving, in that much of their processing power derives *not* from the *vector* and *settling* operations that characterize the "new" paradigm (involving weight and activation vectors), but from cooperative data exchanges between separate modules.

I return to these issues presently, but let us first consider a different sort of objection that may arise. It might be argued that the "challenge" I pose above is not especially worrisome for the connectionist. After all, it is well known that connectionist networks often display emergent side effects. Furthermore, we know that some networks trained via backpropagation have already displayed some degree of compositionality, as evidenced in the capacity of the St. John & McClelland network (1990) to assign correct semantic interpretations to novel sentences. In reply, it should be noted that the degree of *skill* compositionality, required to solve the transverse patterning task, is of a radically different kind than any compositionality displayed by networks that assign semantic representations to novel sentences. The degree of semantic compositionality displayed by Hadley & Hayward's (1997) Hebbian network is markedly greater than that evidenced by St. John and McClelland, but even the Hadley-Hayward network displays no compositionality of entirely separate skills.

Indeed, I know of no non-modular connectionist network, whether eliminativist or not, which exhibits skill compositionality remotely approaching the level required in the transverse patterning problem. Admittedly, in the case of the Marcus generalization task (*a blicket is a ...*), it may not be obvious at first blush that humans employ multiple, *separate* prior skills to solve the task, but we know that, at the very least, a capacity to understand a range of natural language is presupposed. Moreover, *this* capacity is so complex and multi-faceted, that any number of competent linguists would affirm that a variety of separate skills are involved.[4]

Returning now to earlier comments on a *modular* approach to skill compositionality, I would stress that, in my view, such an approach is promising. Indeed, I have argued in (Hadley, 1999) that whenever a variety of markedly distinct skills are involved in a task (such as the skills I have noted above), it is likely that separate modules are involved. Such modules may very well be spatially distributed, and subject to some degree of noisy interactions with other modules, but for computational reasons they should still be regarded as distinct modules. However, in that same paper, I argued that humans are demonstrably able to employ their skill modules in *novel combinations*. To place the issue in a very small nutshell, the mere fact that humans can follow specific kinds of novel rules, within mere seconds after being told such a rule, suffices to show that the brain can transfer information (or data) along sets of *combinatorially adequate pathways* between the separate skill modules. I argued further, by an examination of logically possible cases, that the existence of such combinatorial pathways entails that at least one of several types of classically recognized architectures is present in the human cognitive system. Space limits do not permit a detailed recapitulation of these arguments. However, I submit that a *modular* approach to achieving the impressive "side effects" noted in the "hard cases" which have concerned us, does not represent the type of solution that would appeal to researchers who view connectionism as a radically new paradigm. In any case, neither the hybrid-modular approach, nor the single-holistic-network approach has yet been shown to yield side-effects even approaching those involved in the transverse patterning example.

volved no task possessing an underlying structure identical to that of task T.

[4]Examples of such separate skills include: (1) the ability to recognize distinct words, (2) the ability to recognize highly ungrammatical sentences, (3) the ability to form the past tense of verbs.

References

Boden, M. & Niklasson, L. (2000). Semantic systematicity and context in connectionist networks. *Connection Science*, 12, 111-142.

Fodor, J.A. & Pylyshyn, Z.W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition*, 28, 3-71.

Hadley, R.F. (1993). Connectionism, explicit rules, and symbolic manipulation. *Minds and machines*, 3, 183–200.

Hadley, R.F. (1994a). Systematicity in connectionist language learning. *Mind and Language*, 9, 247-272.

Hadley, R.F. (1994b). Systematicity revisited: reply to Christiansen and Chater and Niklasson and van Gelder. *Mind and Language*, 9, 431-444.

Hadley, R.F. & Hayward, M.B. (1997). Strong semantic systematicity from Hebbian connectionist learning, *Minds and Machines*, 7, 1-37.

Hadley, R.F. (1999). Connectionism and novel combinations of skills: implications for cognitive architecture. *Minds and Machines*, 9, 197-221.

Hadley, R.F., Rotaru-Varga, A., Arnold, D.V., & Cardei, V.C. (to appear). Syntactic systematicity arising from semantic predictions in a Hebbian-competitive network, *Connection Science*.

Marcus, G. (1998). Rethinking eliminative connectionism. *Cognitive Psychology*, Vol. 37.

Marcus, G. (in press). *The Algebraic Mind*. (Cambridge, MA: MIT Press).

Phillips, S. (2000). Constituent similarity and systematicity: the limits of first-order connectionism. *Connection Science*, 12, 1-19.

Niklasson, L.F. and van Gelder, T. (1994). On being systematically connectionist. *Mind and Language*, 9, 288-302.

St. John, M.F. and McClelland, J.L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, 46, 217-257.

Similarity: A Transformational Approach

Ulrike Hahn (HahnU@Cardiff.ac.uk)
School of Psychology, Cardiff University,
PO Box 901, Cardiff, CF10 3YG, UK.

Lucy B. C. Richardson (RichardsonLB@Cardiff.ac.uk)
School of Psychology, Cardiff University,
PO Box 901, Cardiff, CF10 3YG, UK.

Nick Chater (Nick.Chater@Warwick.ac.uk)
Institute for Applied Cognitive Science, Department of Psychology,
University of Warwick, Coventry, CV4 7AL, UK.

Abstract

Representational Distortion is a new account of similarity in which the transformation distance between representations determines similarity: entities that are perceived to be similar have representations that are readily transformed into one another, whereas dissimilar entities require numerous transformations. Here we present experimental evidence in favour of this viewpoint.

Introduction

The breadth of cognitive and social contexts in which similarity is invoked as an explanatory construct is vast. Similarity forms part of the explanations of memory retrieval, categorization, visual search, problem solving, learning, linguistic knowledge and processing, reasoning, and social judgment. The two classical approaches to similarity are the spatial account (e.g., Nosofsky, 1986), which represents similarity in terms of distance in psychological space, and Tversky's (1977) contrast model which views similarity as a function of common and distinctive features of the entities under comparison. Both of these accounts have been used successfully in cognitive modeling, however both also suffer from fundamental limitations (see Hahn & Chater, 1997, 1998a, 1998b). These models are restricted in scope by the fact that they define similarity over very specific and simple types of representation: points in space or feature sets. However, the representation of complex real-world stimuli, from faces to auditory scenes, is typically assumed to require *structured* representations, that can explicitly describe objects, their parts, properties and the relations between them. Relational information of this kind cannot readily be encoded using lists of features or dimensional values (Hahn & Chater, 1998a).

The present paper considers a recent theoretical approach to similarity, Representational Distortion (henceforth, RD; Hahn & Chater, 1997; Chater & Hahn, 1997), which aims to provide a theoretical framework applying to similarity judgements. According to RD, the similarity between two entities is a function of the "complexity" required to "distort" or "transform" the representation of one into the representation of the other. The simpler the necessary transformation, the more similar they are assumed to be.

How can the complexity of the transformation between two representations be measured? At a theoretical level, Hahn and Chater draw on a branch of mathematics, Kolmogorov complexity theory (Li & Vitanyi, 1997) that provides a rigorous and general way of measuring the complexity of representations and transformations between them. In intuitive terms, according to Kolmogorov complexity theory, the complexity of a representation is the length of the shortest computer program that can generate that representation. The idea is that representations that can be generated by a short program are simple; those that require longer programs are complex. We will not consider the virtues of this measure of complexity here, except to note that it supports substantial applications in the cognitive and computing sciences (Chater, 1999).

Kolmogorov complexity has a natural application as a measure of similarity between representations. The simplest measure is the length of the shortest program that "distorts" one representation into the other. According to this viewpoint, the degree to which two representations are similar is determined by how many instructions must be followed to transform one into the other. For example, the conditional Kolmogorov complexity between the sequences 1 2 3 4 5 and 2 3 4 5 6 is small, because the simple instructions add 1 to each digit and subtract 1 from each digit suffice to transform one into the other. In the same way, 1 2 3 4 5 and 2 4 6 8 10 (multiply/divide each digit by 2) are presumed to be similar. On the other hand, 1 2 3 4 5 and 3 5 7 9 11 are viewed as less similar, because two operations are required to transform one into the other (e.g., multiply by 2 and add 1). Finally, two entirely unrelated representations will be maximally dissimilar because there will be no efficient way of transforming

one representation into the other. In this case, the most efficient transformation will involve deleting the first representation, and reconstructing the second from scratch, because there is no shared information between the objects that can be exploited. RD should be viewed as a general framework for understanding similarity, rather than as a specific cognitive account in competition with the spatial or featural views. RD can capture these accounts as special cases (see Chater & Hahn, 1997 for derivations) and thus does not contrast but rather subsumes these accounts. Another motivation for RD is that it aims to provide an explanation for the utility of similarity in inference, for example, to categorize items on the basis of the categories of similar items. To build a concrete psychological account of similarity we need to consider (i) the nature of the mental representations that are relevant to making a similarity judgement; (ii) the set of transformations or instructions that can be used to distort one representation into another; (iii) any constraints on the ability of the cognitive system to discover simple transformations between mental representations.

Despite its generality, RD makes clear empirical predictions. First and foremost, is the prediction that transformations are relevant to similarity. It is this prediction for which the current paper provides empirical support. Crucially, though our own interest in establishing the relevance of transformations to similarity judgments is driven by our research program on RD, the relevance of the general issue of similarity and transformations, and thus of our results, extends beyond our particular theory. As we will demonstrate, the experimental evidence presented here raises substantial problems for classical theories of similarity and raises novel issues for any future work on similarity.

Previous work

The central claim of RD, that similarity is based on transformation distance, has several tentative precursors in the experimental and computational literature. The two most directly relevant experimental studies are by Imai (1977) and Franks and Bransford (1975). Imai proposed that pattern similarity between strings of either filled or unfilled circles was based on transformational structure. He found support for this claim in terms of a qualitative relationship between the number of transformations between two patterns and their judged similarity. However, no statistical analysis was performed. Franks and Bransford (1975) sought to extend Posner and Keele's (1968) work on prototype abstraction, replacing the original random dot patterns with simple geometric figures. Underlying the stimulus set was a prototype that was not shown during training; all other items in the stimulus set were derived from this prototype through the application of one or more simple transformations. Recognition ratings were directly related to transformational distance to the prototype, with the prototype itself receiving the highest rating. Finally, the account has some resonance in the perception literature where transformational explanations have been used to explain figural regularity or "goodness", as well as figuring in theories of object recognition.

In summary, there is currently no clear experimental evidence for the importance of transformations in the context of similarity, despite previous research hinting at this idea. Three experiments were designed to address this issue. Each of the experiments shared the same basic correlational design and differed only in their stimulus materials. The aim was to establish transformational distance as a predictor of perceived similarity, while at the same time providing evidence for the limitations of featural (and spatial) accounts.

Experiment 1

Experiment 1 was based on Imai (1977), and uses sequences of filled or unfilled circles. Transformational distance was manipulated in terms of the number of operations such as mirror imaging, reversal, phase shift, insertion and deletion that were necessary to convert the test stimulus into the target. This is best illustrated with an example stimulus pair, shown below.

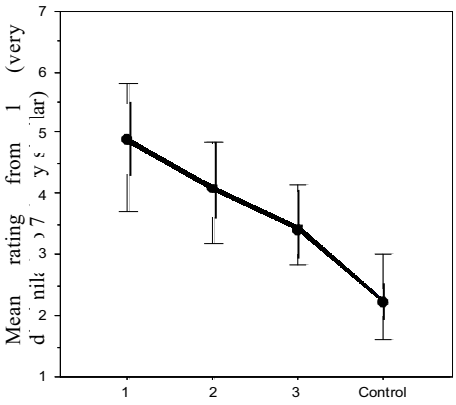
The two rows of "blob" patterns can be transformed into one another through the application of a single operation that "reflects" one row to create the other row in a mirror image. This prediction contrasts with that of the featural account (Tversky, 1977), if we adopt the natural assumption that features correspond to individual blobs. According to this viewpoint, blob patterns should be similar to the extent that they overlap.

Participants 35 undergraduates psychology students.
Materials The stimuli consisted of strings of black and white blobs, presented in pairs. The key transformations were phasic, reversal and mirror, with the addition of insertions and deletions. Any one of these in isolation constituted a single transformation. There were 16 examples of single transformations in this experiment, (4 each of phasic, reversal, mirror and deletion). Multiple transformations were achieved by combining two or more of the above operations. The total set of 56 comparison pairs consisted of 16 examples of two transformational changes (four each of reversal & mirror, reversal & phasic, deletion & mirror and insertion & phasic) and 16 examples of three transformational changes (four each of deletion, reversal & mirror, deletion, reversal & phasic, insertion, reversal & mirror and insertion, reversal & phasic). As a control, there were also 8 pairs of stimuli that were unrelated (or so multiply transformed as to make the transformations unperceivable). Each pair of stimuli was printed horizontally onto a single sheet of paper

together with brief instructions and a rating scale from 1 (very dissimilar) to 7 (very similar). These sheets were then placed into a different random order for each participant and bound into a booklet.

Results

Bivariate correlations between number of transformations and mean similarity rating of each item were found to be highly significant with Spearman's rho = -.69, $p < .005$. The comparison featural model which left aligned the two rows of blobs and counted the number of (mis)matching features fared considerably worse: Spearman's rho = -.28, $p < .05$. Analysis of individual subject ratings confirmed these findings, revealing great conformity across participants: 25 of 35 participants showed significant correlations as predicted. Such consistency was not found using the featural model, with only 8 of the 35 participants showing a significant correlation. The general relationship between number of transformations and mean similarity ratings is graphed below. The results suggest, somewhat surprisingly (see e.g. Shepard, 1987), an approximately linear relationship.



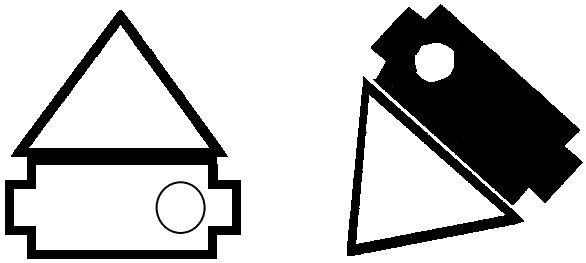
No. of transformations

Discussion

Experiment 1 provides evidence of a statistically significant relationship between transformation distance and perceived similarity to complement Imai's (1977) more qualitative data. The results of Experiment 1 thus confirm both Imai's original intuition and the predictions of the RD framework. In direct comparison, the featural model fares considerably worse in predicting participants' ratings. Consequently the results also provide evidence for limitations in featural approaches. It is, of course, possible that more powerful featural descriptions of the data could be found, but, at present, none are available. Crucially, any putative featural explanation of this kind requires an independent motivation of the features adopted, that is, the postulated features must themselves not be motivated exclusively by salient transformations. Otherwise, the featural description becomes an entirely redundant mimicry. The materials of Experiment 2 make this point more clearly.

Experiment 2

This experiment used simple geometric shapes related by different transformations, e.g., the pair of items shown below has a transformation distance of two as they can be made identical through rotation and color change of one object part. Here, there is no obvious way to apply a featural model for contrast purposes. Furthermore, many of the "features" such as the orientation of an item in a pair where one has been rotated are salient only because of the relevant transformations. This means that central object "features" will be derivative on the transformations present: e.g., orientation is unlikely to have cognitive salience in a comparison until orientation is manipulated through rotations. Consequently, though it might, in principle, be possible to derive featural descriptions for our stimulus items, these descriptions would be likely to implicitly underscore the importance of transformations, rather than providing an alternative to relying on transformations. As in Experiment 1, the prediction is that number of transformations will be negatively correlated with degree of perceived similarity.

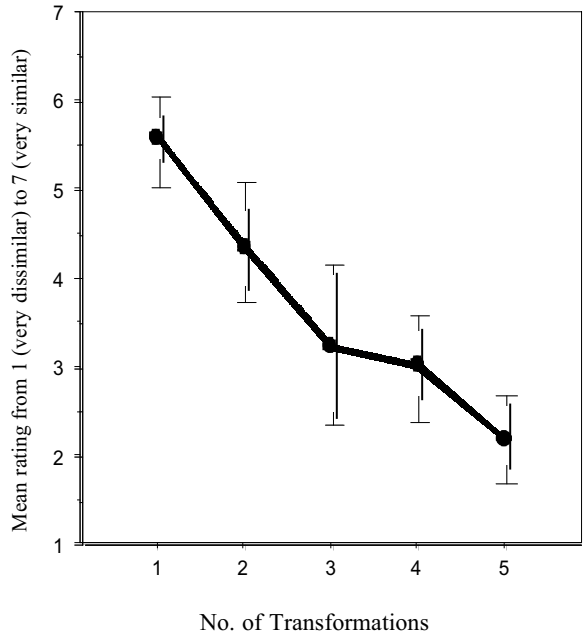


Participants 21 psychology undergraduates.

Materials There were three alternative "target" geometric shape line drawings. The construction of the stimulus sets began with the target shapes to which a single transformation was then applied. Examples of single adjustments were stretching the whole object or changing all striped areas to filled-in black areas. Multiple transformations were constructed by using a combination of such techniques, one at a time. For each of the three target stimuli there four examples each of one, two, three, four and five transformations. This made a set of 20 pairs of pictures for each base target, yielding 60 in total. Each transformed geometric shape was printed onto a separate page together with its corresponding "target", which was always placed to the right of the transformed item. At the top of each page were a set of instructions and a rating scale from 1 (very dissimilar) to 7 (very similar).

Results

Bivariate correlations between number of transformations and mean similarity rating of each item were highly significant with Spearman's rho = -.89, p = .000. Analysis of individual participants' ratings again revealed great consistency across subjects, with 19 of the 21 participants showing a significant correlation as predicted. The relationship between number of transformations and mean similarity ratings as graphed below closely matches that found in Experiment 1. The relationship between transformation distance and similarity is again approximately linear.



Discussion

Experiment 2 provide further supports the role of transformations in the context of similarity, mirroring the results of Experiment 1 with very different stimulus materials. The materials of Experiment 2 also illustrate how even natural "features", such as orientation are influenced by transformations. Many of the very "features" that a featural account might posit seem salient due to the transformational relationships between the two compared objects. This is indicative of the general bi-directional relationship posited by RD theory between object representation and transformation, with perceived transformations influencing which aspects of an object become salient and vice versa.

Experiment 3

Experiment 3 sought to take the argument against featural and spatial representations one step further, by using materials for which such representation schemes are obviously inadequate, because they depend on relational structure. We used 3D objects assembled from (typically) three Lego bricks: one large brick, colored blue; a medium size yellow brick; and a small red brick. Each similarity comparison comprised two objects assembled from these three bricks, albeit in different spatial arrangements. Despite the extreme simplicity of these stimuli, relational information (i.e., information about relative position, such as, for example, that the red brick as on top of the yellow brick) is paramount to the representation of the composite objects. However, the appeal of these materials is not limited to the difficulties they pose for featural or spatial accounts. From a transformational perspective, the Lego brick objects are of interest for two reasons. First, they allow an initial examination of the role of transformations in the similarity assessment of real-world objects, albeit maximally simple ones. Second, these materials support a whole new range of transformations to complement those investigated in Experiments 1 and 2. Our assumption, here, was that the judged similarity between pairs of objects would be determined primarily by the physical manipulations required to turn a target object into the comparison object.

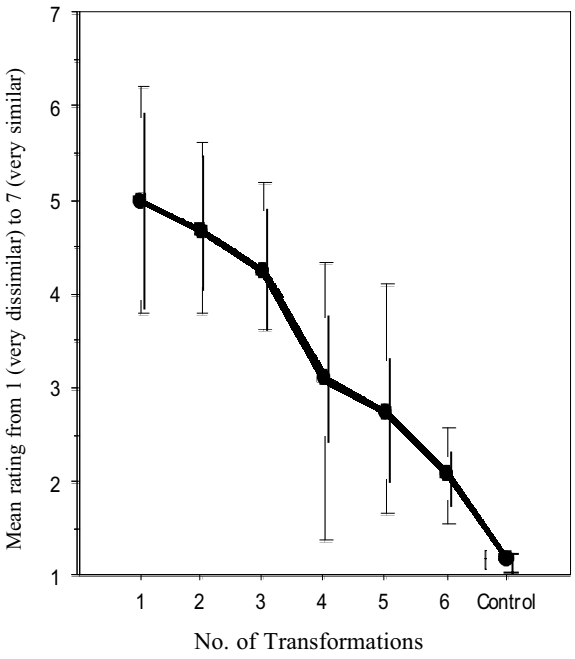
Participants 27 psychology undergraduates.

Materials The stimuli were based upon an initial "target" array of three Lego bricks, (a four-point square red brick, upon a six-point oblong yellow brick, on top of an eight-point oblong blue brick) arranged into a particular three-dimensional structure. Apart from the two examples that were chosen to be totally unrelated to the target, all of the Lego stimuli were constructed by transforming the original target object a set number of times, prior to the experiment. The researcher began with the target arrangement and made adjustments to it, each of which constituted one transformation. For example, one adjustment (or transformation) could involve moving an object within the arrangement, substituting a brick for one of a different size or colour, adding an additional brick, subtracting an existing brick, or changing the order of the bricks within the arrangement. The transformations required to create each arrangement were counted and the entire set of stimuli constructed so that it comprised of 42 items (10 examples of one and 10 of two transformations, and five examples each of three, four, five and six transformations away from the "target". In addition there were two items that were unrelated--multiply transformed--stimuli). Once these arrangements of Lego bricks had been constructed, each was glued into a permanent structure.

Procedure Participants were shown the "target" Lego brick object. They rated how similar they perceived each stimulus to be to the target, on a scale of 1 (very dissimilar) to 7 (very similar). Every participant rated all the Lego stimuli within the set of 42 items.

Results

Bivariate correlations between number of transformations and mean similarity rating of each item where again highly significant, Spearman's rho = -.76, $p < .005$. Analysis of individual participants' ratings again revealed great consistency across subjects, with all 35 exhibiting a significant correlation between number of transformations and rated similarity. The general relationship between number of transformations and mean similarity ratings (shown below) is again very similar to that that found in Experiments 1 and 2.



Discussion

Despite the very different materials and set of relevant transformations, the results of Experiment 3 closely match those of Experiments 1 and 2. Again, they provide evidence for the importance of transformations in explaining similarity judgements, and are difficult to account for in terms of the featural or spatial views of similarity, which cannot easily handle relational information. Thus, these results are in line with the predictions of a central tenet of the RD account: that the transformational relationship between representations of two objects determines their judged similarity. But the results Experiment 3 also have broader implications. The inherently relational nature of the materials in Experiment 3 poses a problem for any representational scheme which does not allow structured representations; conversely it lends support to any account such as structural alignment theories to which such structured representations are central. Similarly, the result that number of transformations is a significant predictor of perceived similarity lends support to the general notion of an influence of transformations on similarity, whether or not the particular framework of RD theory is adopted.

General Discussion

The results of all three experiments provide robust evidence as to the importance of transformations in explaining similarity judgments, across a variety of different stimulus types. These findings support the central tenet of the RD theory of similarity, that similarity is based on the complexity of the transformation between the representation of two items. These results also provide new evidence for the limitations of classical accounts of similarity. All three experiments provide evidence against purely featural views of similarity. Experiment 1 provides a direct test. The version of the featural model we tested (assuming that features correspond to blobs) is not the most sophisticated featural description possible, given that, in principle, any property including all higher-order regularities such as "symmetry" etc. could be posited as features (Tversky, 1977). Crucially, however, a more sophisticated featural account which succeeds in providing comparable or even superior data fits must not only first be found, it must also be independently motivated. Given that theories can be stretched beyond all recognition through the addition of suitable post hoc auxiliary assumptions, a crucial factor in evaluating competing accounts must not only be whether an account can be made compatible with a particular pattern of data, but also whether it in any way predicted it.

In Experiment 1, there is nothing in featural theories of similarity that would naturally give rise to the predictions made on the basis of transformations in this experiment. The sequences of filled circles lend themselves naturally to a featural decomposition on a one by one basis due to the fact that the "object" is readily parsed into a set of individual circles. Many of the relevant "features" of the geometric shape stimuli in Experiment 2 become cognitively salient only through transformational contrast between the two comparison objects (for example, the feature "orientation" highlighted by the transformation "rotation"). Consequently, transformations are explanatorily prior. The use of simple formations of Lego bricks in Experiment 3 demonstrates the central representational weakness of featural accounts - their inability to deal with structured representations and thus adequately represent relational information. The challenge presented by Experiment 3 is to identify even a remotely suitable featural description, given the inherent relational nature of materials and transformations. The limitations of featural accounts exposed by this series of experiments is equally shared by spatial models of similarity, whether they are based on multi-dimensional scaling or standard connectionist networks. It must again be stressed that from the perspective of RD theory, featural and spatial accounts of similarity are not

wrong, they are simply too restricted to cope with the flexibility of transformations available to the cognitive system.

Interestingly, issues related to this research have been raised in philosophy. Goldman (1986) suggests that the lawfulness of human similarity judgments might be furthered by an inherent preference ranking for transformations, which comes in to play where multiple transformational sequences could link the same stimulus pair. This question links closely with a central issue for future research, that of the relative "cost" or "weight" of individual transformations. Single transformations need not be equal in cost or 'effort'. Such inequalities arise automatically in the theory of Kolmogorov complexity, the general mathematical framework on which RD theory draws. Here, deletions, for example, tend to be less costly than insertions, because deletions only require a specification sufficient to identify the component for deletion, whereas insertions require a complete specification of the additional component. What weightings of this kind, if any, are intrinsic to the cognitive system is an issue we are currently seeking to determine through the investigation of perceived similarity for different single transformations. Information as to relative transformational costs will be crucial for more detailed cognitive modeling and thus constitute a major issue for future research. Another potentially interesting area is to apply the approach to different domains; particularly those that appear to require structured representations where RD can be utilized in a straightforward way. For example, two postures of a hand, in terms of a specification of joint angles can be compared simply. Given the transformations likely to be salient in cognitive processing involving motor control, we might expect that 'similar' hand positions would correspond to positions that can readily be transformed into each other.

We have presented a new account of similarity, Representational Distortion, according to which the judged similarity between a pair of items is determined by the complexity of the transformation between the mental representations of those items. We have tested the central tenet of the account in three experiments, finding that transformational complexity is, indeed, inversely related to similarity. These results present a challenge for other accounts of similarity, based on feature comparison or spatial distance; and they indicate that the view that similarity can be explained in terms of transformation merits further theoretical and empirical investigation.

Acknowledgements

This work was funded by the Leverhulme Trust, and NC is supported by EU Grant RTN-HPRN-CT-1999-00065.

References

Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *Quarterly Journal of Experimental Psychology, 52A*, 273-302.

Chater, N. & Hahn, U. (1997) Representational Distortion, Similarity, and the Universal Law of Generalization. In, *Proceedings of the Interdisciplinary Workshop on Similarity and Categorization*, SimCat97, University of Edinburgh. Dept. of Artificial Intelligence, University of Edinburgh.

Franks, J.J. & Bransford, J.D. (1975) Abstraction of Visual Patterns. *Journal of Experimental Psychology: Human Perception and Performance, 4*, 65-74.

Goldman, A.I. (1986) *Epistemology and Cognition*. Cambridge:MA.Harvard University Press.

Hahn, U. & Chater, N. (1997). Concepts and similarity. In K. Lamberts and D. Shanks (Eds.). *Knowledge, concepts and categories*, (pp. 43-92). Hove, England: Psychology Press.

Hahn, U. & Chater, N. (1998a). Similarity and rules. Distinct? Exhaustive? Empirically distinguishable? *Cognition, 65*, 197-230.

Hahn, U. and Chater, N. (1998b) Understanding Similarity: a Joint Project for Psychology, Case-Based Reasoning and Law. *Artificial Intelligence Review, 12*, 393-427.

Imai, S. (1977) Pattern Similarity and Cognitive Transformations. *Acta Psychologica, 41*, 433-447.

Li, M. & Vitanyi, P. (1997). *An introduction to Kolmogorov complexity and its applications* (2nd edition). New York: Springer-Verlag.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39-57.

Posner, M. I. & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353-363.

Shepard, R. N. (1987) Toward a universal law of generalization for psychological science. *Science, 237*, 1317-1323.

Tversky, A. (1977) Features of similarity. *Psychological Review, 84*, 327-352.

A Parser for Harmonic Context-Free Grammars

John Hale (hale@cogsci.jhu.edu)
Department of Cognitive Science
The Johns Hopkins University
3400 North Charles Street; Baltimore MD 21218-2685
Paul Smolensky (smolensky@cogsci.jhu.edu)
Department of Cognitive Science
The Johns Hopkins University
3400 North Charles Street; Baltimore MD 21218-2685

Abstract

Harmonic Grammar is a connectionist-derived grammar formalism, of which Optimality Theory is a kind of limiting case. Harmonic Grammar is expressive enough to specify the trees that are correct parses on a given context-free grammar. Here, we show how to construct a connectionist parsing network which finds correct parses given a sentence, of if none exist, signals a rejection. Finally, a brief comparison to other connectionist parsing work is provided.

Harmonic grammar is a grammar formalism which uses only soft rules of the following form:

> If condition X is violated, then the well-formedness (Harmony) of the structure is diminished by C_X.
>
> [Legendre et al., 1990, 388]

A linguistic theory in Harmonic grammar is a set of soft rules and a set of representational possibilities. Applying the soft rules to any representation yields the Harmony value for that representation. A representation with maximal Harmony from some class defined by common substructure — the inputs — is said to be the winning candidate. In Optimality Theory [Prince and Smolensky, 1993] the soft rules are ranked and the C_X values are arranged so that no number of violations of lower-ranked soft rules could ever outweigh a violation by a higher ranked one.

Harmonic grammar is expressive enough to specify the context-free languages [Smolensky, 1993]. Here Harmony maximization is made to serve as a processing algorithm for a parser. The design for the parser is a way of systematically arranging networks of threshold logic units so that they implement exactly the Harmonic grammar rules for context-free grammars. They maximize the Harmony of representations that share the same terminal entries. The general idea is that these networks do this by de-activating pieces of bad analyses until only correct analyses remain.

How Harmonic Grammar specifies context-free languages

Harmonic context-free grammar rules (as first presented in [Smolensky, 1993]) are integer occurrence and co-occurrence penalties that are defined on trees. Trees

Figure 1: A legal derivation?

whose Harmony value is zero are successful derivations which prove that their yield is generated on the grammar. All others have negative Harmony which indicates that the yield is not in the language generated by the grammar.

Derived from connectionist principles, Harmonic grammar assumes the following form for the Harmony function.

$$H(\mathbf{a}) = \sum_{\alpha<\beta} a_\alpha W_{\alpha\beta} a_\beta \quad \sum_{\alpha} a_\alpha b_\alpha \tag{1}$$

In this case, **a** is the representation of a parse tree as a vector. W is a symmetric weight matrix, and H is a sum of terms containing pairs of elements from the vector **a**. As a consequence of this form of the Harmony function, the conditions X in the soft rules are restricted to referring to at most two structures; Harmony maximization is quadratic optimization.

This is an apparent problem for phrase-structure grammar, since no pairwise check of any two symbols from the tree depicted in figure 1 on the grammar $\begin{matrix} X & AP \\ X & QB \end{matrix}$ could reveal that the tree is not a valid derivation of AB from X. It would seem that to check a rule with two children, rules that refer to three pieces of the representation at once are needed, implying a cubic Harmony function. But if this is so, then surely to check a rule with three children would require a quartic Harmony function. Rather than adopting Harmony functions of higher and higher degree, Harmonic context-grammars are defined for context-free grammars[1] in a special normal form where pairwise evaluation is sufficient to check global wellformedness: Harmonic Normal Form.

[1] Throughout, V is the set of all grammar symbols, Σ is the subset of V which are terminals, R is the set of rules or productions represented as (symbol,string) pairs, and S is the distinguished start symbol. See [Lewis and Papadimitriou, 1981] for notation, definitions and fundamental results on context-free grammars.

Definition 1 (branchingrhs) *Let $G = (V, \Sigma, R, S)$ be a context-free grammar in Chomsky Normal Form, A $V - \Sigma$ a nonterminal from G and γ a string in V . Then* branchingrhs$(A) =$ A γ $R :$ $\gamma > 1$

Definition 2 (Unique Branching) *A context-free grammar $G = (V, \Sigma, R, S)$ satisfies the Unique Branching condition if, for all nonterminals A $V - \Sigma$,* branchingrhs(A) 1.

Unique branching insists that for every parent, at most one ordered pair of children is licensed by the grammar. This is the condition that defines Harmonic Normal Form and makes pairwise evaluation sufficient to specify context-free grammar trees. For example, the apparent problem mentioned previously would be solved if only the following grammar, which satisfies the Unique Branching condition, could be used instead.

X	$X\ 1$
X	$X\ 2$
$X\ 1$	AP
$X\ 2$	QB

On this grammar, the tree in figure 1 is assigned negative Harmony. If X had been expanded by the first rule, the parent would be $X\ 1$ and the tree would be penalized for lacking P. If the parent were $X\ 2$ the tree would be penalized for lacking Q. The extra nonterminal encodes which original context-free rule was used, but this contextual information is not needed at higher levels of the parse tree, and the unary rules helpfully remove it.

When the grammar satisfies the Unique Branching condition, a natural interaction between Harmonic grammar rules becomes sufficient to evaluate local trees. Because H adds up harmony penalties, the grammar effectively computes an "AND" at the site of each bracketed parent. All that remains is to specially balance the soft rule weights so that pairs in local trees licensed by the grammar exactly balance out to 0 Harmony and those in ill-formed local trees receive some kind of harmony penalty, ultimately leading to $H < 0$ for the whole tree. A set of rules that does this, G_H, is given below.

G_{HNF}	G_H
a	R_a: If a is at a node, add -1 to H
A	R_A: If A is at a node, add -2 to H
$A\ i$	$R_{A\ i}$: If $A\ i$ is at a node, add -3 to H
start symbol S	R_{root}: If S is at the root, add 1 to H
A α ($\alpha = a$ or $A\ i$)	If α is a left child of A, then add 2 to H
$A\ i$ BC	If B is a left child of $A\ i$, add 2 to H If C is a right child of $A\ i$, add 2 to H

[Smolensky and Legendre, 2001, chapter 10]

Grammar preprocessing

The penalties that figure into the Harmonic grammar rules are going to be connection weights and unit biases in a neural network that parses the grammar. The relation between the grammar and the neural network is established by two grammar transformations. The first ensures that the Unique Branching condition is upheld.

Definition 3 (HNF transform) *Let $G = (V, \Sigma, R, S)$ be a context-free grammar in Chomsky Normal Form and let A, B, C, X $V - \Sigma$. The HNF transform HNF of G is a new grammar $HNF(G) = (V\ , \Sigma, R\ , S)$ where for each nonterminal A that appears in i branching rules of the form A BC, each such rule is replaced by two new rules containing a new nonterminal not in $V - \Sigma$, having the forms A $A\ i$ and $A\ i$ BC. Call the set of new nonterminals that appear in these additional rules* bracket$(V\ )$. *The transformed set V is the union of* bracket$(V\ )$, *the old nonterminals X $V - \Sigma$ such that* branchingrhs$(X) = 0$ *and the old terminals Σ.*

If a symbol is an element of the set bracket$(V\ )$ it is called "bracketed" otherwise it is "unbracketed."

The second transformation adds information about string positions to every rule, and restricts the grammar to only describing sentences of a certain maximum length. Since this maximum can be arbitrary large, it seems reasonable to maintain that context-free grammars for infinite languages are described in the limit [Charniak and Santos, 1987].

The annotation of string positions enables grammar symbols to directly serve as parser items. An item B_{jm} is an assertion about the input string that means "there is a constituent of type B that spans sentence positions j to m."

Definition 4 (Itemification of a binary rule) *The itemification of a binary context-free rule A BC to a sentence length is the set of rules given by the schema A_{jkm} $B_{jk}C_{km}$ for all $j, k, m = 0\ldots$ such that $j < k$ m.*

Definition 5 (Itemification of a unary rule) *The itemification of a unary context-free rule A B to a sentence length is the set of rules given by the schema A_{jm} B_{jkm} for all $j, k, m = 0\ldots$ such that $j < k$ m.*

Grammars resulting from both transformations include complex symbols of the form $A\ i\ _{jkm}$. These symbols express the assertion that there is an A-type constituent spanning sentence positions j to m which was derived via the i^{th} A-rule, and the left child's yield stops at position k. In this way, k plays the role of a back-pointer that addresses a bracketed parent's children.

Definition 6 (Itemification of a grammar) *Let $G = (V, \Sigma, R, S)$ be a context-free grammar in Chomsky Normal Form. The itemification of G carried out for a sentence length is a grammar $ITEM(G,\ ) = (V\ , \Sigma\ , R\ , S\ )$ in which*

1. Σ *contains* -1 *symbols labeled $v_{jj\ 1}$ (where $j = 0\ldots\ -1$) for each terminal symbol v in Σ.*
2. S *is a new start symbol labeled S_0*
3. R *contains the itemification of each rule r in R.*

and V consists of all the symbols appearing in S , R , Σ .

Itemification ensures that children are directly adjoining, and in the right order. For example, (neglecting k's for a moment) if the grammar contains $X_{13} \quad Y_{12}\ Z_{23}$ it will definitely not contain $X_{13} \quad Y_{13}\ Z_{23}$ where a part of Z's yield — the symbol from position 2 to position 3 — is enveloped by Y's yield.

It is also convenient to define the width of two-index itemified symbols X_{ij} from V as $width(X) = j - i$. Further, we suggest (without going into the proofs here[2]) that there are cover homomorphisms between proper parse relations on each of $ITEM(G, \ )$ and $HNF(G)$ and G. In the case of $ITEM(G)$ this homomorphism is only defined for parses of sentences of length . For these sentences, call these homomorphisms f_{ITEM} and f_{HNF}. These are the "inverse mappings" that, given a parse on a transformed grammar, supply a parse on the untransformed grammar — essentially undoing the work of their namesakes. The basic idea is that both transformations only add or rename rules, rather than deleteing them.

Finally, define the parent set of a grammar symbol to be the set of all nonterminals that appear on the left-hand side in rules that involve the symbol in question on the right-hand side.

Definition 7 (Parent set) *Let* $G = (V, \Sigma, R, S)$ *be a context-free grammar and* $\gamma_0, \gamma_1 \quad V$ *Then the set of all possible parents of an element* $X \quad V - \Sigma$ *is* $\text{parents}(X, G) = P : \ \gamma_0, \gamma_1 \quad V$ *such that* $P \quad \gamma_0 X \gamma_1 \quad R$

Every context-free grammar with a finite number of rules has a "maximal parent multiplicity" $p_{\max(G)}$, the highest number of possible parents for any symbol. All of these concepts and definitions will be used to completely specify the parsing network in the next section.

Hopfield network

The parsing network is a Hopfield network with units whose states take on just the values 0 and 1. The network shall be constructed to parse the grammar $ITEM(HNF(G), \ ) = (V, \Sigma, R, S)$ with maximal parent multiplicity $p_{\max} = p_{\max(ITEM(HNF(G), \ ))}$. There are $\alpha = 1 \ldots V$ threshold logic units which update themselves according to the transition rule

$$a_\alpha \quad \begin{array}{l} 0 \text{ if } \sum_{\alpha=\beta} W_{\alpha\beta} a_\beta \quad b_\alpha < 0 \\ 1 \text{ if } \sum_{\alpha=\beta} W_{\alpha\beta} a_\beta \quad b_\alpha \quad 0 \end{array}$$

where $W_{\alpha\beta}$ are connection weights and b_α are biases. Let f_α denote the application of the transition rule to the α^{th} threshold logic unit. Then a network update f_{network} is defined by $f_{\rho(1)} \quad f_{\rho(2)} \quad \ldots \quad f_{\rho(V)}$ where ρ indexes the entries in a random permutation of $1 \ldots V$.

[2] See [Nijholt, 1980, chapter 2] for discussion of grammar covers.

Entries in the undirected $V \times V$ weight matrix W are indexed by grammar symbol and are only nonzero if one indexed symbol is bracketed and the other is unbracketed. Without loss of generality, identify the α^{th} grammar symbol as the bracketed one $(A\ i_{jkm})$ and the β^{th}, as the unbracketed one (A_{jm}).

If there is a binary rule $\alpha \quad \beta\gamma$ or $\alpha \quad \gamma\beta$ the weight between units α and β is 1.

If there is a unary rule $\alpha \quad \beta$ then the weight between units α and β is the same as the maximal parent multiplicity, $p_{\max}$.

Otherwise the weight is zero.

The vector of V biases, **b** is all negative. Component b_α is set to one of three possible values.

If α is an unbracketed start symbol of $width =$ then $b_\alpha = -1$, or, if α is not a start symbol,

if α is bracketed then $b_\alpha = -(p_{\max} \quad 2)$, or else

α is unbracketed and $b_\alpha = -(p_{\max} \quad 1)$

These weights and biases reflect the kinds of input that bracketed and unbracketed units need to be correctly supported by units representing parents above them and units representing children below them.

Bracketed units By the Unique Branching condition, these units have indegree three. They receive input from exactly one (unbracketed) parent and exactly two (unbracketed) children. If all three of these neighbors are in the 1 state, then the net input is $p_{\max} \quad 1 \quad 1$ representing, respectively, the contributions from the parent and each child. This exactly balances the bias $-(p_{\max} \quad 1)$ and keeps the bracketed unit in that state, always updating via the 0 transition. Bracketed units that are on are guaranteed to have correct parents and children on.

Unbracketed units These units can be connected to as many as $p_{\max}$ (bracketed) parents. Even in the worst case, in which all possible parents are in the 1 state, an unbracketed units' $-(p_{\max} \quad 1)$ bias makes sure that it can only be in the 1 state itself when supported by at least one unbracketed child. Unbracketed units that are on are guaranteed to have at least one correct child on.

By construction, W is symmetric (if two symbols are in a parent-child relationship they are also in a child-parent relationship) and has zeros along the diagonal (no two symbols are in dominance relationships with themselves), making the results on convergence of Hopfield networks [Hopfield, 1982] applicable.

During the network's operation, only the states of units associated with symbols of width > 1 may change. Width-1 units, specifying the input to be parsed, are

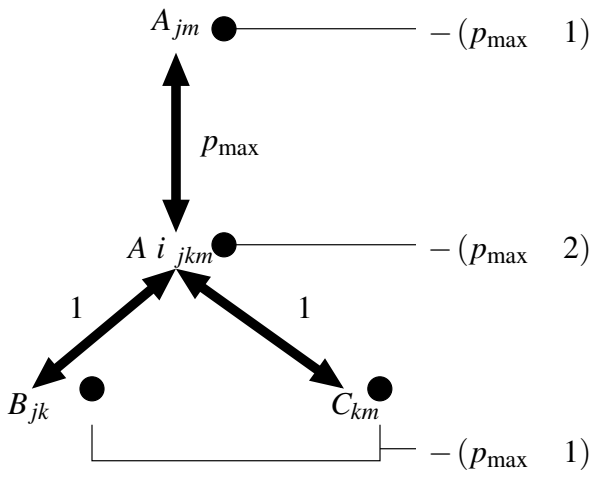


Figure 2: Fundamental parsing network block

clamped. For simple asynchronous updating, the change in a unit a_α's activation is

$$\Delta a_\alpha = \begin{array}{rl} 1 & \text{if the old state was 0 and} \sum_\beta W_{\alpha\beta} a_\beta \quad b_\alpha \quad 0 \\ 0 & \text{if the old state was 1 and} \sum_\beta W_{\alpha\beta} a_\beta \quad b_\alpha \quad 0 \\ 0 & \text{if the old state was 0 and} \sum_\beta W_{\alpha\beta} a_\beta \quad b_\alpha < 0 \\ -1 & \text{if the old state was 1 and} \sum_\beta W_{\alpha\beta} a_\beta \quad b_\alpha < 0 \end{array}$$

The corresponding change in Harmony is

$$\Delta H = \Delta a_\alpha \quad \sum_{\alpha=\beta} W_{\alpha\beta} a_\beta \quad b_\alpha$$

Since Δa_α is positive when $\left(\sum_\beta W_{\alpha\beta} a_\beta \quad b_\alpha\right)$ is, and Δa_α is negative when $\left(\sum_\beta W_{\alpha\beta} a_\beta \quad b_\alpha\right)$ is, H is increasing whenever $\Delta a_\alpha = 0$. But H is also clearly bounded from above, at least by $\sum_\alpha b_\alpha \qquad \sum_\alpha \sum_{\beta>\alpha} W_{\alpha\beta}$, and so cannot increase indefinitely. Therefore the dynamics reaches a maximum, at which point $\Delta H = 0$. At this point $f_{\text{network}}(\mathbf{a}^{\text{stable}}) = \mathbf{a}^{\text{stable}}$.

Note that unbracketed units will only turn off if all units representing their (bracketed) child-options are off. Bracketed units will switch off if any of their neighbors switch off. The basic arrangement repeated throughout the network is depicted in figure 2.

Theorem 1 (Correctness) *Let* $\mathbf{a}^{\text{stable}}$ *be a stable state of a Hopfield network constructed as above to parse* $ITEM(HNF(G), \ ) = (V, \Sigma, R, S)$ *whose initial state* $\mathbf{a}^0$ *is determined by the input string* $v = v_{01} v_{12} \ldots v_{\ -1}$ *in the following way:*

If v_{ij} *is contained in the input and the* m^{th} *grammar symbol is* v_{ij}*, then the* m^{th} *component of the initial state is 1.*

If v_{ij} *is not contained in the input and the* m^{th} *grammar symbol is* v_{ij}*, then the* m^{th} *component of the initial state is 0.*

Otherwise the m^{th} *component of the initial state is 1.*

Then, if the final state is 1 for a unit associated with a start symbol of $\text{width}(\alpha_m) = \ $ *, then the set* $\alpha_m : a_m = 1$ *determine a shared packed forest [Tomita, 1986] of v-parses on G. Otherwise the parser has rejected v.*

Proof: We must show that if a unit representing a start symbol spanning the entire input is in the 1 state at $\mathbf{a}^{\text{stable}}$, then all trees determined by sequences of choices about which activated bracketed child-units to move to from activated unbracketed parent-units, going from the root to the leaves, are correct parses of v.

If a unit representing a start symbol spanning the entire input is in the 1 state, it must be because its -1 bias has been counterbalanced by activation from at least one child, since by definition there are no parents for start symbols.

Select one of these bracketed children that are also in the 1 state. As bracketed units, being on implies a full and correct set of neighbors in the 1 state. Two of these neighbors are bracketed children.

Continue the proof by selecting arbitrarily from among the activated bracketed children at each successive unbracketed unit. This selected unit must be part of a correct parse in virtue of a grammar rule, or it would not be activated. Eventually because the network is finite this selecting and traversing must end at clamped, unbracketed units of width 1.

Each selection of a bracketed unit from the perspective of an unbracketed parent is an unpacking of one choice that has been packed in the shared-packed parse forest. The representation is shared because no symbol is represented more than once.

Since all of the units that are on are part of some correct parse corresponding to some sequence of bracketed-rule selections, for each such correct parse there must be a sequence $\pi = \pi_0, \pi_1, \ldots, \pi_n$ of rules which each describe one piece of local tree structure. Since the above argument did not depend on which bracketed-rule unit was selected at each point, all sequences of selections result in correct parses and all the resulting π stand in a proper parse relation with v on $ITEM(HNF(G), \ )$. The proper parse relation on G is $f_{HNF} \quad f_{ITEM}(v, \pi)$.

Corollary 1 (Completeness & the initial state) *If the initial state* $\mathbf{a}^0$ *includes enough* $a_j = 1$ *to describe a parse of w then that parse will be represented in the final state.*

Proof: The parser's operation can only switch bracketed units $\alpha \quad \text{bracket}(V)$ in the 1 state into the 0 state, and not the other way around, because W is constructed so that α's row, W_α has exactly three nonzero entries, and their sum is $(p_{\max} \quad 2)$. By construction these nonzero entries are at exactly the columns for the two unique children and unique parent. α's bias has also been constructed to be exactly $-(p_{\max} \quad 2)$. So given that α is on, it must be that all of α's neighbors are on and that they are licensed by the grammar. But bracketed units in the

1 state with *correct* parents and children do not change their state. So if all correct parents and children from a parse are present in the initial state, and no bracketed units can switch off, then all correct parents and children must still be on in the stable state.

Example

As an example, consider the ambiguous grammar $S \rightarrow AB$, $A \rightarrow AA$, where S is the start symbol. We follow [Nijholt, 1990] in assuming that preterminal rules such as $A \rightarrow$ v don't play a role and parsing may begin at the nonterminals. The input sequence $AAAB$ is ambiguous on this grammar between an analysis where the left most pair of A's form a constituent $(((AA)A)B)$ and one where the middle two A's form one $((A(AA))B)$. Either analysis ultimately will be compatible with a correct parse.

To build a parser for this grammar for sentences of length $= 4$, the grammar is first transformed by HNF. Even though the grammar was already in Harmonic Normal Form, HNF here serves to explicitly encode the unary/binary status of each rule.

$$\begin{array}{ll} A\ 1 & \rightarrow AA \\ A & \rightarrow A\ 1 \\ S\ 1 & \rightarrow AB \\ S & \rightarrow S\ 1 \end{array}$$

Itemification is then performed, resulting in a larger grammar in which all possible rule applications have been annotated with string position indices for every possible location at which they could be applied.

$$\begin{array}{ll} A\ 1\ 012 & \rightarrow A01\,A12 \\ A\ 1\ 013 & \rightarrow A01\,A13 \\ A\ 1\ 023 & \rightarrow A02\,A23 \\ \vdots \quad \vdots & \quad \vdots \\ A01 & \rightarrow A\ 1\ 011 \\ A02 & \rightarrow A\ 1\ 012 \\ \vdots \quad \vdots & \quad \vdots \\ S\ 1\ 012 & \rightarrow A01\,B12 \\ S\ 1\ 013 & \rightarrow A01\,B13 \\ \vdots \quad \vdots & \quad \vdots \\ S01 & \rightarrow S\ 1\ 011 \\ S02 & \rightarrow S\ 1\ 012 \\ S03 & \rightarrow S\ 1\ 013 \\ S03 & \rightarrow S\ 1\ 023 \\ \vdots \quad \vdots & \quad \vdots \end{array}$$

There are 54 units, each associated with symbol in this transformed grammar. The network runs until it reaches a stable state, at every transition increasing Harmony. The Harmony values for one simulation run are shown in figure 3.

In the final state, units representing the symbols $S04$, $S\ 1\ 034$, $A03$, $A\ 1\ 013$, $A\ 1\ 023$, $A02$, $A\ 1\ 012$, $A\ 1\ 123$, $A13$, $A01$, $A12$, $A23$ and $B34$ are all in the one state, and all others are in the zero state. Since the start symbol

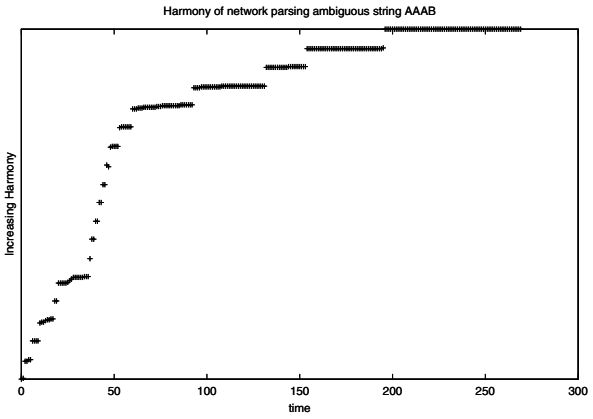


Figure 3: Operation of parsing network on example grammar

$S04$ is activated, we can interpret the parse as having accepted $AAAB$. To determine the parses, we conceptually traverse downward from $S04$ to $S\ 1\ 034$, and then to $A03$ and $B34$. From $A03$, there is a choice of which of the two ambiguous parses to be taken. Both are represented by activated units all of which are part of correct parses that figure into a shared packed parse forest. Selecting $A\ 1\ 013$ determines one parse, and selecting $A\ 1\ 012$ determines the other, just as in chart parsing.

Comparison

One difference between the architecture of this Harmonic grammar parser and various other deterministic connectionist parsers ([Fanty, 1985], [Nijholt, 1990], [Sikkel, 1997]) resides in the lack of central control over evaluation order. The formulation here is in terms of fixed points for randomly-ordered, Harmony-increasing updates. Despite this apparent freedom, the de-activation of unsupported units proceeds bottom-up, an order effect which follows from the connectivity of the network.

As in treatments that avoid Harmony minima through simulated annealing ([Selman, 1985], [Howells, 1988]) the parser's progress can be tracked by examining the current value of H, although here the parser state is not probabilistic.

Other comparisons invite exciting extensions. The work of Hopfield [Hopfield, 1984] suggests that the results for linear threshold units should extend straightforwardly to more realistic neural models, while that of Stolcke [Stolcke, 1989] points the way to more linguistically realistic unification-based grammars.

Perhaps the most intriguing comparison is to Optimality Theory itself. As in Optimality Theory, where all representational possibilities are said to come from *Gen*, the Hopfield network parser described here starts from a state in which all possible constituents are represented. As processing progresses, units representing constituents that lack support given the input string deactivate themselves. In this way the parser acts as a filter

that removes ungrammatical analyses from a initial universe of conceivable analyses. The parser is implementing constraints from a *Con* that contains the soft rules G_H. However, because constraint interaction is numerical, strict domination does not necessarily hold: two or more violations of R_a can be just as bad, or worse, than a single violation of R_A even though $R_A \quad R_a$.

Conclusion

In fact, the ultimate goal of the larger research program of which this work forms a part is the integration of insights from three different sources: formal grammar, constraint-based processing, and linguistic theory. Harmonic grammar is a competence theory that can declaratively specify context-free and other formal languages. Here we have shown that a simple performance theory can be constructed that incorporates this competence theory in a relatively straightforward way into a procedural specification for parsing using abstract neural computing units. In the overall program, however, the role of formal languages is to benchmark theories of human parsing. The analogies to OT in this simple performance theory suggest that such an architecture may be flexible enough to accommodate insights into human language processing from OT syntax and constraint-based approaches to psycholinguistics.

References

[Charniak and Santos, 1987] Charniak, E. and Santos, E. (1987). A connectionist context-free parser which is not context-free, but then it is not really connectionist either. In *Proceedings of the 9th Annual Conference of the Cognitive Science Society*, pages 70–77, Hillsdale, NJ. Erlbaum.

[Fanty, 1985] Fanty, M. (1985). Context-free parsing in connectionist networks. Technical Report TR147, Rochester Computer Science Department.

[Hopfield, 1982] Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States*, 79:2554–2558.

[Hopfield, 1984] Hopfield, J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences of the United States*, 81:3088–3092.

[Howells, 1988] Howells, T. (1988). VITAL: a connectionist parser. In *Proceedings of 10th Annual Meeting of the Cognitive Science Society*, pages 18–25.

[Legendre et al., 1990] Legendre, G., Miyata, Y., and Smolensky, P. (1990). Harmonic grammar – a formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pages 388–395, Cambridge MA. Erlbaum.

[Lewis and Papadimitriou, 1981] Lewis, H. R. and Papadimitriou, C. H. (1981). *Elements of the Theory of Computation*. Prentice-Hall, Englewood Cliffs, NJ.

[Nijholt, 1980] Nijholt, A. (1980). *Context-Free Grammars: Covers, Normal Forms and Parsing*. Number 93 in Lecture Notes in Computer Science. Springer-Verlag, Berlin.

[Nijholt, 1990] Nijholt, A. (1990). Meta-parsing in neural networks. In Trappl, R., editor, *Proceedings of the 10th European Meeting on Cybernetics and Systems Research*, pages 969–971, Teaneck NJ. World Scientific.

[Prince and Smolensky, 1993] Prince, A. and Smolensky, P. (1993). *Optimality theory: constraint interaction in generative grammar*. MIT Press. Forthcoming.

[Selman, 1985] Selman, B. (1985). Rule-based processing in a connectionist system for natural language understanding. Technical Report CSRI-168, University of Toronto Computer Science Department.

[Sikkel, 1997] Sikkel, K. (1997). *Parsing Schemata: a framework for specification and analysis of parsing algorithms*. EATCS Texts in Theoretical Computer Science. Springer.

[Smolensky, 1993] Smolensky, P. (1993). Harmonic grammars for formal languages. *Advances in neural information processing systems*, 5:847–854.

[Smolensky and Legendre, 2001] Smolensky, P. and Legendre, G. (2001). *Architecture of the Mind/Brain: neural computation, optimality and universal grammar in cognitive science*. Forthcoming.

[Stolcke, 1989] Stolcke, A. (1989). Unification as constraint satisfaction in structured connectionist networks. *Neural Computation*, 1:559–567.

[Tomita, 1986] Tomita, M. (1986). *Efficient parsing for natural language: a fast algorithm for practical systems*. Kluwer Academic Publishers, Boston.

Models of Ontogenetic Development for Autonomous Adaptive Systems

Derek Harter (dharter@memphis.edu)
Department of Mathematical Sciences; University of Memphis
Memphis, TN 38152 USA

Robert Kozma (rkozma@memphis.edu)
Department of Mathematical Sciences; University of Memphis
Memphis, TN 38152 USA
Arthur C. Graesser (a-graesser@memphis.edu)
Department of Psychology; University of Memphis
Memphis, TN 38152 USA

Abstract

Biological organisms display an amazing ability during their ontogenetic development to adaptively develop solutions to the various problems of survival that their environments present to them. Dynamical and embodied models of cognition (Clark, 1997; Edelman & Tononi, 2000; Franklin, 1995; Freeman, 1999a, 1999b; Freeman & Kozma, 2000; Freeman, Kozma, & Werbos, 2000; Hendriks-Jansen, 1996; Kelso, 1995; Kozma & Freeman, 2001; Port & van Gelder, 1995; Skarda & Freeman, 1987; Thelen & Smith, 1994) are beginning to offer new insights into how the numerous, heterogeneous elements of neural structures may self-organize during the development of the organism in order to effectively form adaptive categories and increasingly sophisticated skills, strategies and goals. In this paper we present models of ontogenetic development built on neurologically inspired, bottom-up, dynamic approaches to embodied category formation such as those done by Freeman (1975, 1999b), Freeman and Kozma (2000), Kozma and Freeman (2001), Verschure and Voegtlin (1999) and Edelman (1987), Edelman and Tononi (2000). We believe that building on such mechanisms from an embodied dynamical perspective will produce autonomous agents that display greatly increased flexibility in their behavior. Such models will represent a better understanding of how the brains of biological organisms not only form perceptual categories of their environments during development, but also develop effective patterns of behavior through the dynamic self-organization of neurological patterns of activity.

Introduction

Biological organisms develop effective behaviors simply by perceiving and acting upon their environment in real time. Their learning is always guided by their basic needs. Through their experience with the environment, they begin to embody, anticipate and exploit the regularities of their ecological niche in the service of their intrinsic needs. Some models of learning and development for autonomous systems are beginning to display some of these properties. (Almássy, Edelman, & Sporns, 1998; Edelman et al., 1992; Freeman & Kozma, 2000; Kozma & Freeman, 2001; Verschure, Kröse, & Pfeifer, 1992; Verschure, Wray, Sporns, Tononi, & Edelman, 1995) These abilities include the formation of embodied, organism significant categories through experience; the development of active searching and anticipation of relevant stimuli; the development of a repertoire of skills, or action loops, for the effective transformation of environmental problems and the exploitation of environmental regularities in the service of intrinsic needs.

In this paper we will present some of the most important properties of dynamical and embodied cognition. We will also discuss the properties of ontogenetic development of skills, strategies and goals in biological organisms that make it a particularly powerful mechanism of learning. We will look at examples of existing systems that display properties of dynamical and embodied cognition. And finally we discuss our own plans for creating models of the ontogenetic development of behavior in autonomous adaptive systems.

Embodied Cognition

Embodied cognition is an emerging viewpoint in cognitive science that emphasizes many differing aspects from the standard cognitive hypothesis (Clark, 1997; Hendriks-Jansen, 1996; Pfeifer & Scheier, 1998). In the standard view of cognition, the mind is the product of the manipulation of symbolic representations of the problem in order to produce solutions and generate intelligent behavior (Johnson-Laird, 1988; Newell & Simon, 1972, 1976; Newell, 1990). The environment is perceived and transduced into symbolic representations. These symbols encode the current state of the environment and the problem to be solved. They can be manipulated, independent of the environment, to discover solutions to the problem and produce intelligent behavior for the organism.

In an embodied view of cognition, intelligence in biological organisms does not arise through the static manipulation of amodal symbols and representations. Instead, organisms are seen to be embedded in their environments in fundamental ways. Through their real time experiences with their bodies and environments, they begin to embody the salient aspects of situations in ways that guide future perception and behavior towards improved performance. Experience with their ecological niche develops expectations of the environmental regularities that are of benefit to the intrinsic needs and desires of the organism. The organism actively learns to seek out expected stimuli that are relevant to the desires and needs of the organism at a particular moment.

There are many concepts associated with an embodied perspective of cognition. We will briefly present some of

the more important concepts in the next sections.

Embodied Organisms are Complete Organisms

Biological organisms are currently the only examples capable of producing a full range of intelligent, adaptive behavior. Standard views of cognition place no special emphasis on the fact that these natural examples of cognition are **complete** organisms. In the standard view of cognition, it seems plausible that by connecting together many specialized subsystems that solve problems in limited, specialized domains, eventually a complete intelligence will be produced.

From an embodied perspective, we are not likely to understand natural cognition from such a piecemeal approach to studying and building systems. Instead, we must examine and build complete cognitive systems. In this context, complete refers to systems that are autonomous and adaptive. Autonomous systems are those that have certain intrinsic needs, and that are able to produce behavior that is capable of satisfying those needs consistently over time. Pfeifer (Pfeifer & Scheier, 1998) characterizes autonomy as the ability of the organism to maintain its critical, intrinsic values within a zone of viability. This is often referred to as "homeostasis". Adaptivity refers to organisms that are capable of modifying their behavior so that they can more efficiently maintain their critical parameters in their zones of viability.

Studying complete cognitive systems is important for several reasons. Classical approaches to modeling cognition often tackle toy problems in limited domains. The hope is that the techniques developed can then be scaled up to the full problems of cognition. This approach to studying cognition has failed to produce clear insights into how such methods could eventually be scaled up. Embodied cognition, with its emphasis on complete systems, maintains that the answer is not to start with toy environments. Instead we should begin by studying simple, but complete, organisms, in more realistic environments (Brooks, 1990; Pfeifer & Scheier, 1998). Only complete organisms are capable of developing embodied representations and displaying intentional behavior.

Active, Action-Oriented Representations

Another important difference of embodied and classical perspectives concerns the nature of the representations developed and used by the organism. In a classical perspective, symbols are seen as passive structures that are syntactically manipulated to produce solutions. In an embodied perspective, representations are much more intimately tied to the intrinsic needs of the organism. Clark (1997) calls such structures *action-oriented representations*. Action-oriented representations are not passive representations of the state of the environment as it exists at some time. They are continuously updated from sensory information, and they continuously prescribe possibilities for action. Gibson (1979) has called this the concept of affordances, where the representations *afford* opportunities for action for the organism.

The World Represents Itself

Classical models of cognition often experience an exponential explosion of computational power as the environment increases in complexity. An embodied approach to cognition avoids this problem because it advocates the use of simple, cheap, action-oriented representations. From an embodied perspective, it is better to use cheap and active sensing to inform oneself of the state of the environment, rather than building complex representations of the environment. Brooks (1995) states this principle as "the world is its own best model". Embodied cognition avoids the use of costly and detailed representations. Cheap, quick, active, specialized sensing of the environment is preferred. Instead of maintaining a complex representation of the state of the environment, we simply direct specialized sensory apparatus to directly perceive the information required for behavior. This approach helps keep the need for computation from exploding in complex environments.

Emergence of Solutions through Collective Activity

A key concept of embodied cognition is the emergence of solutions from many parallel, distributed activities. In an embodied perspective, intelligence is seen as emerging from the parallel activity of many cooperating and competing processes. As in connectionist models, parallel emergence of solutions provides many benefits to the behavior of the system. Such emergent solutions are robust and resistant to damage; tolerant of noisy, incomplete data; satisfy general goals and yet are variable and context dependent. They are also fast, able to produce solutions easily in real time demanding environments. Unlike most classical connectionist modeling, embodied cognition views recurrent, non-linear interactions as a crucial property in the emergence of solutions.

Developing Within the Environment

The emergence of solutions through many parallel processes is not simply a product of the non-linear interactions of components in the organism's brain. Intelligent behavior also emerges as the product of the interaction of simple behaviors with a complex environment. Simple, instinctive behaviors are seen as intelligent when they are coupled with local environmental cues (Braitenberg, 1984). Development of action-oriented representations aids in this process. Organisms learn simple actions that, when coupled with appropriate learned stimuli, yield intelligent, purposeful behavior.

Clark (1997) says that embodied minds use extensive external scaffolding. The ecological niche of the organism provides many consistent cues for intelligent behavior. Most intelligent behavior in natural organisms involves the fast recognition and exploitation of such opportunities, not in complex planning and reasoning. Also, most organisms tend to offload complex planning and reasoning tasks onto the environment. They do this by allowing the state of the environment to represent the

progression of the problem solving task. One example, given by Rumelhart, McClelland, and The PDP Research Group (1986), is in the behavior of people when multiplying large numbers. Most people can instantly recognize and produce the answer to simple, single digit multiplication problems, of the type 7 x 7 = 49. However, when given the task of multiplying large numbers together, say 4356 X 1897, they invariably resort to pencil and paper, or even a calculator. People do not compute large chains of complicated reasoning and logic. Instead they offload the representation of the progress of the task onto the environment by maintaining the state of the problem solving task with environmental cues. In this case, people make marks on paper (the environment) to keep track of their problem solving progress, while reducing the problems to those simple ones that they can directly recognize and solve. Embodied cognition sees this type of external scaffolding not as simply useful, but as a prevalent and pervasive method used by cognitive systems to reduce computational complexity and perform problem solving tasks in real time.

Better Imperfect than Late

Biological cognition is exemplified by fast pattern completion. It has evolved to produce behavior in real time. The behavior does not necessarily have to be perfect, so long as it is good enough for the continued survival of the organism (at least until the next crisis occurs). Organisms are continually presented with threats and dangers that must be handled immediately in order to ensure their survival. Such requirements do not favor solutions that take large amounts of time. Natural cognition seems to be built upon a foundation of fast pattern recognition and behavior generation keyed to threats and opportunities for action. The embodied cognitive viewpoint recognizes this fundamental feature of natural cognitive systems. According to Port and van Gelder:

> ”The cognitive system is not a discrete sequential manipulator of static representational structures; rather, it is a structure of mutually and simultaneously influencing *change*. Its processes do not take place in the arbitrary, discrete time of computer steps; rather, they unfold in the *real* time of ongoing change in the environment, the body, and the nervous system. (Port & van Gelder, 1995, pg. 3)”

The Dynamics of Development

The ontogenetic development of behavior provides a powerful mechanisms by which organisms learn to organize effective patterns of behavior for performing the necessary tasks of survival. There are many properties of this type of development. It is fundamentally a self-organizing process, in which the constraints of body and environment guide the system towards discovering certain patterns of behavior. Development of behavior in organisms is not so much a process of finding complex chains of effective behaviors, but in finding salient perceptual cues and effective manipulations that simplify and transform the task environment into problems that are directly recognizable and solvable. Problem solving in natural cognitive systems is more often the application of many transformations until the problem is sufficiently simplified to be directly solved. Clark (1997) calls such phenomena action loops. Kirsh and Maglio (1994) call actions that are primarily performed to transform and simplify the task environment epistemic actions.

Problem solving behavior in biological organisms does not tend to be encoded as static, procedural steps. Instead, organisms develop a wide repertoire of action loops and epistemic actions. Development of behavior takes the form of learning more and better action loops for the effective manipulation and transformation of problems. As an organisms repertoire of action loops grows, they become better able to deal with a wide variety of subtle differences in the problems they need to solve. Their solutions become both robust and efficient with experience in problem solving in the environment.

Development of Embodied Cognition

Thelen and Smith (1994), Thelen (1995) envision the development of behavior in cognitive systems as an ontogenetic landscape of stable and unstable attractors and repellors. As the body of the organism changes, new opportunities for behavior are created and destroyed. Development is seen as a reduction of the degrees of freedom of the system as useful patterns for solving problems are discovered. As stable solutions to problems develop, these in turn change the ontogenetic landscape, opening up new opportunities for some behaviors, and closing off opportunities for others. Development is the discovery of stable patterns of behavior, given the current constraints of the body and the environment.

Natural cognitive systems display both physical and behavioral development. Physical changes in a maturing organism are continually reshaping the ontogenetic landscape, destabilizing previously stable solutions, and forcing the system into finding new patterns of behavior. Natural cognitive systems also display this flexibility in the development of behavior for problem solving. Sequences of behaviors are not learned so much as behaviors that change the state of the environment and thus cue the next behavior in the sequence.

Self-Organization of Behavior

Theories of the self-organization of patterns in nonequilibrium systems provide new insights into the creativity and flexibility displayed by biological organisms (Kelso, 1995). Many of the desirable properties of development in biological organisms make sense only in view of nonlinear dynamics. According to Kelso:

> “The thesis here is that the human brain is *fundamentally* a pattern-forming self-organized system governed by nonlinear dynamical laws. Rather than compute, our brain dwells (at least for short times) in metastable states: it is poised on the brink of instability where it can switch flexibly and quickly.

> By living near criticality, the brain is able to anticipate the future, not simply react to the present. (Kelso, 1995, pg. 26)"

The development of problem solving behavior in biological organisms displays these important properties. Solutions are developed that are flexible, efficient and quick. Such systems are not simply reactive, they learn to anticipate and actively seek out future stimuli.

Bottom Up Neurological Models of Categorization and Action

Some systems have been developed that display properties of dynamic and embodied cognition as discussed above. In this section we present four interesting examples of research that display dynamic, self-organizing category formation and development of behavior. These are all examples of systems that have been built using neurologically inspired, intermediate level neural dynamics.

Distributed Adaptive Control

Distributed Adaptive Control, or DAC (Pfeifer & Verschure, 1992; Pfeifer & Scheier, 1998; Verschure et al., 1992; Verschure & Voegtlin, 1999) is an example of a model of learning based on large scale neural dynamics. At its heart, DAC is a model of classical conditioning, or the learned association of a response to a conditioned stimuli. In the DAC model, there are three levels of control: reactive, adaptive and reflective control.

The reactive level is prewired in the model, and represents the intrinsic values of the autonomous agent. In the case of DAC, the robot instinctively turns away from things when it bumps into them. This represents the value of avoiding damage from collisions with the environment. In addition to a collision sensor, a special sensor for target acquisition is present. DAC is hardwired to move towards the target when it is detected by the target sensor.

The next level is the adaptive control layer. In this layer representations of the states of long range sensors are slowly associated with events that happen in the reactive control layer. So, for example, the system will learn to avoid collisions by associating the profiles of objects sensed with the long range sensor to collisions and the subsequent activation of avoidance behavior. DAC is also capable of learning and exploiting the regularities of the ecological niche it finds itself in. So, if targets are always found behind openings in walls, DAC is capable of learning this association and begins to search out such openings since they tend to lead to finding the targets in the environment.

The final layer of DAC is the Reflective control layer. At this level sequences of actions are formed and remembered through developing sequential representations. This level represents the addition of long term memory to the basic mechanisms of adaptive learning.

DARWIN

DARWIN (Almássy et al., 1998; Edelman, 1987; Edelman et al., 1992; Edelman & Tononi, 2000; Sporns, Almássy, & Edelman, 1999; Verschure et al., 1995) is another neurologically inspired model that is capable of learning and developing representations simply by interacting within its environment. At the heart of Edelman's DARWIN systems is the classification couple. In a classification couple, two maps of neuronal groups receive input from separate sensors. The two maps are wired together with many reentrant connections. As a result of reentrant coupling and the change of synaptic strengths, corresponding classification patterns begin to be associated and mutually activate one another in the maps. Thus, for example, the feel (tactile map) and shape (visual map) of an object become functionally correlated through repeated experience with the objects in the environment. The correlated patterns of activity in the maps represent coordinated properties of objects encountered within the environment.

DARWIN III is capable of self-organizing categories of objects that it encounters in its environment, and of learning appropriate behavior patterns. DARWIN is capable of learning to track moving objects in its environment and also of directing its manipulator in a targeted manner in order to manipulate its environment. DARWIN III is also capable of adaptive learning of behavior, like DAC. It learns to associate visual properties of desirable and undesirable objects, to the feel of the object. As it gains experience in the environment, it no longer needs to touch a bad object in order to avoid it. It has formed associations between the visual and tactile maps, and it begins to avoid undesirable objects upon seeing them.

KIII: Mesoscopic Dynamics

The discovery that brain dynamics operate in chaotic domains has profound implications for the study of higher brain function (Skarda & Freeman, 1987). A chaotic system has the capacity to create novel and unexpected patterns of activity. It can jump instantly from one mode of behavior to another, which manifests the fact that it has a collection of attractors, each with its basin, and that it can move from one to another in an itinerant trajectory. It retains in its pathway across its basins a history, which fades into its past, just as its predictability into its future decreases. Transitions between chaotic states constitute the dynamics that we need to understand how brains perform such remarkable feats as abstraction of the essentials of figures from complex, unknown and unpredictable backgrounds, generalization over examples of recurring objects never twice appearing the same, reliable assignment to classes that lead to appropriate actions, and constant up-dating by learning.

The KIII model (Freeman & Kozma, 2000; Kozma & Freeman, 2001) consists of various sub-units; i.e., the KO, KI, and KII sets. The KO set is a basic processing unit, and its dynamics is described by a 2nd order ordinary differential equation. By coupling a number of excitatory and inhibitory KO sets, KI(e) and KI(i) sets

are formed. Interaction of interconnected KI(e) and KI(i) sets forms the KII unit. Examples of KII sets in the olfactory system are the olfactory bulb, anterior olfactory nucleus and prepyriform cortex. Coupling KII sets with feed-forward and feedback connections, one arrives at the KIII system.

KIII shows very good performance in learning input data and it can generalize efficiently in various classification problems. KIII has a high dimensional chaotic attractor in the basal state. It can be destabilized by sensory stimuli and switched to a lower dimensional attractor wing that represents a previously learned memory pattern.

Basic Intentional System: The Limbic System

We consider biological organisms to be behaving intelligently when they act in ways that will enhance their current and future survival. The behavior exhibited by biological organisms is often very creative and flexible. Yet such behavior is always directed towards the satisfaction of the basic needs of the organism. Freeman (1999a, 1999b) describes such behavior as intentional behavior. Intentionality provides a key concept that links the neurodynamics of brains to goal-directed behavior.

One of the primary acts of intentional behavior is in directing sensory observation in expectation of information to guide future actions. Both the formation of expectations and the real time dynamic interaction of the organism with the environment are important principles of intentional behavior. Freeman's view of the mechanisms of intentionality is one of nonlinear dynamic interaction of heterogeneous neural elements on many levels and time scales. The neurodynamic architecture of the brain forms many recurrent loops between brain and brain, brain and body, and organism and environment. But the basic architecture of intentional behavior can be found in the simplest and phylogenetically oldest parts of biological brains: the limbic system.

Conclusion and Future Directions

In this paper we have presented an overview of the dynamical and embodied cognitive hypothesis. We have also given an overview of some systems that display category formation and developmental learning of the type we are interested in. We have begun work on our own models of the ontogenetic development of behavior in autonomous systems (Harter & Kozma, 2001a, 2001b). Our own models emphasize the development of action-oriented representations that afford opportunities for action-loop like interactions between the agent and the environment. Such models are based upon the formation of embodied categories from chaotic non-linear dynamics.

We begin with bottom-up neurological models that are capable of chaotic non-linear dynamics (Freeman & Kozma, 2000; Kozma & Freeman, 2001). These neurologically inspired models are neither low nor high level simulations of neurological function, but instead capture behavior of the mesoscopic dynamics of brain function (Freeman & Kozma, 2000). These models of neurological function are capable of the dynamic formation of categories. These dynamic categories can be thought of as models of embodied category formation. We are planning to expand such mechanisms to not only form perceptual categories, but develop and display action-loop like skills in the context of the problem domain. Our goals are to see how far such mechanisms can go in developing problem solving behaviors, and to what extent these behaviors mimic those seen in natural cognitive systems.

Eventually we plan to build simplified models of complete limbic systems. We hope that these models will be capable of displaying forms of true intentional behavior in autonomous adaptive systems. Such models should display some of the characteristic flexibility of the problem solving behavior that develops in natural cognitive agents. We are developing agents in cognitively demanding real time task environments. Beginning with some virtual environments, like the game of Tetris (Kirsh & Maglio, 1992, 1994), we are developing bottom-up neurological models that are capable of category formation and the development of behavior in such environments. We hope to eventually move to more complex environments, and real autonomous robots.

Acknowledgments

Portions of this work were funded by NSF grant SBR-9720314.

References

Almássy, N., Edelman, G. M., & Sporns, O. (1998). Behavioral constraints in the development of neuronal properties: A cortical model embedded in a real world device. *Cerebral Cortex*, *8*, 346-361.

Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology.* Cambridge, MA: The MIT Press.

Brooks, R. A. (1990). Elephants don't play chess. *Robotics and Autonomous Systems*, *6*, 3–15.

Brooks, R. A. (1995). Intelligence without reason. In L. Steels & R. Brooks (Eds.), (pp. 25–81). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Clark, A. (1997). *Being there: Putting brain, body, and world together again.* Cambridge, MA: The MIT Press.

Edelman, G. M. (1987). *Neural darwinism: The theory of neuronal group selection.* New York, NY: Basic Books.

Edelman, G. M., Reeke, G. N., Gall, W. E., Tononi, G., Williams, D., & Sporns, O. (1992). Synthetic neural modeling applied to a real-world artifact. *Proceedings of the National Academy of Science*, *89*, 7267–7271.

Edelman, G. M., & Tononi, G. (2000). *A universe of consciousness: How matter becomes imagination.* New York, NY: Basic Books.

Franklin, S. P. (1995). *Artificial minds.* Cambridge, MA: The MIT Press.

Freeman, W. J. (1975). *Mass action in the nervous system.* New York, NY: Academic Press.

Freeman, W. J. (1999a). Consciousness, intentionality and causality. In R. Núñez & W. J. Freeman (Eds.), (pp. 143–172). Bowling Green, OH: Imprint Academic.

Freeman, W. J. (1999b). *How brains make up their minds.* London: Weidenfeld & Nicolson.

Freeman, W. J., & Kozma, R. (2000). Local-global interactions and the role of mesoscopic (intermediate-range) elements in brain dynamics. *Behavioral and Brain Sciences*, *23*(3), 401.

Freeman, W. J., Kozma, R., & Werbos, P. J. (2000). Biocomplexity: Adaptive behavior in complex stochastic dynamical systems. *BioSystems*.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Houghton Mifflin.

Harter, D., & Kozma, R. (2001a). Ontogenetic development of skills, strategies and goals for autonomously behaving systems. In *Proceedings of the 5th world multi-conference on systemics, cybernetics and informatics (SCI 2001).* Orlando, FL.

Harter, D., & Kozma, R. (2001b). Task environments for the dynamic development of behavior. In *Proceedings of the intelligent systems design and applications 2001 workshop (isda 2001).* San Francisco, CA.

Hendriks-Jansen, H. (1996). *Catching ourselves in the act: Situated activity, interactive emergence, evolution and human thought.* Cambridge, MA: The MIT Press.

Johnson-Laird, P. N. (1988). *The computer and the mind: An introduction to cognitive science.* Cambridge, MA: Harvard University Press.

Kelso, J. A. S. (1995). *Dynamic patterns: The self-organization of brain and behavior.* Cambridge, MA: The MIT Press.

Kirsh, D., & Maglio, P. (1992). Reaction and reflection in tetris. In J. Hendler (Ed.), *Artificial intelligence planning systems: Proceedings of the first annual international conference (aips92).* San Mateo, CA: Morgan Kaufman.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, *18*, 513–549.

Kozma, R., & Freeman, W. J. (2001). Chaotic resonance - methods and applications for robust classification of noisy and variable patterns. *International Journal of Bifurcation and Chaos*, *11*(6).

Newell, A. (1990). *Unified theories of cognition.* Cambridge, MA: Harvard University Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the Association for Computing Machinery*, *19*, 113-126.

Núñez, R., & Freeman, W. J. (Eds.). (1999). *Reclaiming cognition: The primacy of action, intention and emotion.* Bowling Green, OH: Imprint Academic.

Pfeifer, R., & Scheier, C. (1998). *Understanding intelligence.* Cambridge, MA: The MIT Press.

Pfeifer, R., & Verschure, P. F. M. J. (1992). Distributive adaptive control: A paradigm for designing autonomous agents. In F. J. Varela & P. Bourgine (Eds.), (pp. 21–30). Cambridge, MA: The MIT Press.

Port, R. F., & van Gelder, T. (Eds.). (1995). *Mind as motion: Explorations in the dynamics of cognition.* Cambridge, MA: The MIT Press.

Rumelhart, D. E., McClelland, J. L., & The PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge, MA: MIT Press.

Skarda, C. A., & Freeman, W. J. (1987). How brains make chaos in order to make sense of the world. *Behavioral and Brain Sciences*, *10*, 161–195.

Sporns, O., Almássy, N., & Edelman, G. M. (1999). Plasticity in value systems and its role in adaptive behavior. *Adaptive Behavior*, *7*(3-4).

Steels, L., & Brooks, R. (Eds.). (1995). *The artificial life route to artificial intelligence: Building embodied, situated agents.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Thelen, E. (1995). Time-scale dynamics and the development of an embodied cognition. In R. F. Port & T. van Gelder (Eds.), (pp. 69–100). Cambridge, MA: The MIT Press.

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action.* Cambridge, MA: The MIT Press.

Verschure, P. F. M. J., Kröse, B., & Pfeifer, R. (1992). Distributed adaptive control: The self-organization of behavior. *Robotics and Autonomous Systems*, *9*, 181–196.

Verschure, P. F. M. J., & Voegtlin, T. (1999). A bottom-up approach towards the acquisition, retention, and expression of sequential representations: Distributed adaptive control III. *Neural Networks*, *11*, 1531-1549.

Verschure, P. F. M. J., Wray, J., Sporns, O., Tononi, G., & Edelman, G. M. (1995). Multilevel analysis of classical conditioning in a behaving real world artifact. *Robotics and Autonomous Systems*, *16*, 247–265.

Representational Form and Communicative Use

Patrick G.T. Healey (ph@dcs.qmw.ac.uk)
Department of Computer Science;
Queen Mary, University of London, UK.
Nik Swoboda (nswoboda@cs.indiana.edu)
Department of Computer Science;
Indiana University, Bloomington, USA
Ichiro Umata (umata@mic.atr.co.jp)
Yasu Katagiri (katagiri@mic.atr.co.jp)
Media Integration and Communications Laboratories;
ATR International, Kyoto, Japan

Abstract

The form of representations is typically considered to be conditioned by three things: the nature and availability of domain regularities, the perceptual and cognitive abilities of individuals, and the properties of the medium used to construct a representation. This paper reports on an experimental investigation of a fourth constraint on representational form; communicative use. Subjects were given a graphical interaction task in which they produced drawings of pieces of music. The results demonstrate that both level of interaction and communicative context have a marked influence on the form of the representations produced. The results parallel findings for dialogue and indicate that communicative use may be a key constraint on representational form.

Background

In order to be effective in communication, a representation must address, in some reliable way, regularities in the represented world or domain. This intuition naturally suggests a focus on characterising the relationship between the form of representations and the form of a particular domain. For example, between the elements of a picture and the scene or object it represents, or between the structure of a sentence and a mathematical model of the domain. Of course, this relationship may be indirect and arbitrary but where it can be characterised it provides a basis for comparing representations according to properties such as: abstraction, conventionalisation, expressiveness, iconicity, schematisation, and specificity. In cognitive science, three factors are typically cited as moderating these properties. Firstly, the domain to be represented must be sufficiently regular or structured. Secondly, individual perceptual and cognitive limitations constrain both the types of regularity that are identified and the form of the representation used to capture them. Thirdly, representational form is conditioned by the properties of the medium used to produce them.

Significant attention has been directed to analysing the importance of cognitive-perceptual factors and the properties of the medium in conditioning the form of graphical representations (see Scaife & Rogers, 1996, for a review). For example, some authors have highlighted how graphical representations can exploit spatial layout to reduce memory load and facilitate reasoning about a domain (e.g., Stenning & Oberlander, 1995; Shimojima, 1996). The influence of conceptual limitations and perceptual processes in normalising and conventionalising the form of graphical representations has also been investigated (Bartlett, 1932; Tversky, 1981, 1989, 1995). The specific physical properties of graphical media have also been cited as a constraint on representational form. For example, clay discourages the fluent use of detailed graphical forms, favouring simplified, reduced symbols or scripts instead. Historical transitions from pictographic to more abstract scripts have been attributed, in part, to the introduction of clay tablets as the principal writing medium (Tversky, 1995).

These considerations overlook what we take to be the primary function of graphical and linguistic representations: use in communication. Evidence is accumulating to indicate that distinctively interactional or communicative factors constrain the form of descriptions used in dialogue. For example, Garrod and Doherty (1994) showed experimentally that choice between alternative forms of spatial description is primarily influenced by pressure to establish conventions within a sub-community. Importantly, these effects arise independently of domain structure and independently of cognitive-perceptual factors such as individual task expertise (Healey, 1997, 2001). Schwartz (1995) also provides evidence of significant communicative constraints on the form of graphical representations. He studied the differences between the graphical representations produced during problem solving by pairs or individuals. Dyads produced abstract problem representations, such as matrices and graphs, significantly more often than individuals working alone.

This paper reports two experiments that investigate the influence of domain structure, media type, and communicative use on representational form. Our task was designed to meet three basic requirements. Firstly, the precondition that the task domain should exhibit a number of basic regularities which could be captured in a representation. Secondly, that there should be no strong pre-existing representational conventions for carrying out the task. The rationale for this is that our focus on changes in representational form requires a domain for which participants can, in principle, deploy a variety of possible representations and which does not encourage them (or us) to suppose there is a particular 'correct' representation. Existing experimental studies of graphical representation have focussed almost exclusively on tasks and materials, such as maps, Euler circles, circuit diagrams

and program flow charts, which have standard interpretations and which individuals must learn to read in the intended way (cf. Scaife & Rogers, 1996). Thirdly, we wanted a task that is accomplished by exclusively graphical means. This facilitates comparison with other communicative modalities, especially language and is more likely to promote the development of novel conventions. To this end we developed a music communication task. This task involves people producing pictures of pieces of music so that a second person can use them to identify the piece drawn. Music provides a highly structured task domain because of the availability of parameters such as: tempo, intensity, texture, scale and mode. It also provides a domain for which, with the exception of formal notations used by musicians, participants have no pre-existing representational conventions to call on.

Experiment 1

Experiment 1 investigated the effects of domain structure on the form of drawings produced. Musical genre (Jazz vs. Classical) was used as the basic manipulation of domain structure. Two quantitative dependent measures were selected: accuracy of identification of the music from the drawings and time to respond.

Materials 36 pieces of piano solo music were chosen; 18 Jazz pieces and 18 Classical pieces. Each piece was by a different composer or artist and easily recognisable pieces were avoided. The Jazz pieces all included some improvisation and used non-diatonic chord progressions or tones whereas the Classical pieces did not.

All drawing was carried out and captured on a shared virtual whiteboard written specially for the task. The whiteboard was displayed on two LCD tablets (combined graphics tablet and screen) connected to two desktop computers. The whiteboard consisted of a shared drawing area, a strip palette of eight colours and a set of buttons for controlling playback and indicating selections at the top. Subjects could draw using a stylus and lines could be erased by using the reverse end of the stylus.

Subjects 24 participants were recruited from local universities and divided into 12 pairs. They were paid an honorarium for taking part.

Procedure On each trial one participant, the Giver, drew a picture of a target piece of piano music. Givers were free to draw anything they like, subject to the restriction than no letters or numbers should be used. The other member of the pair, the Follower, saw the Giver's drawing developing on the whiteboard and their task was to use this picture to select which of two pieces; the target and a distractor, it corresponded to. Playback and selection of the target and distractor pieces, controlled by buttons at the top of the screen, was self-paced. Followers were asked to make their choice as quickly and as accurately as possible. If a two minute time limit expired before the Follower decided then further drawing by the Giver was blocked and a dialogue window appeared to prompt a final choice. After each trial subjects received feedback about whether the choice was correct. This was repeated for 24 trials with the roles of Giver and Follower alternating between the members of a pair. Music was randomly assigned subject to the constraints that no individual heard the same piece of music twice (as target or distractor), each piece of music occurred equally often as target and distractor, choice of pieces was counterbalanced across pairs and conditions, and order of presentation was randomised for each pair.

Results

Despite some initial hesitation, participants found the task intelligible and engaging. They were able to perform consistently above chance getting, on average, 68% correct ($t_{(44)} = 6.38$, p (two tailed) = 0.00.)[1]. To evaluate effects of the manipulation of genre, two analyses were carried out. Firstly the type of target drawn by the Giver; Classical vs. Jazz. Secondly, the type of discrimination performed by the Follower; same genre (Jazz vs. Jazz and Classical vs. Classical) or different genre (Jazz vs. Classical). The average proportion of correct responses for each pair were analysed in a 2 factor analysis of variance: There was no simple main effect of target type ($F_{(1,11)} = 2.10$, p = 0.17) or discrimination type ($F_{(1,11)} = 2.10$, p = 0.66) and no interaction ($F_{(1,11)} = 0.58$, p = 0.46). The corresponding analysis for average response times also showed no simple main effect of either target type ($F_{(1,11)} = 0.84$, p = 0.38) or discrimination type ($F_{(1,11)} = 2.92$, p = 0.12) and no interaction ($F_{(1,11)} = 2.87$, p = 0.12).

Informal inspection suggested that some of the drawings appeared to be coding emotional affect (e.g., sad vs. happy faces). It was hypothesised that this might reflect the influence of a second aspect of musical form; mode (Major or Minor). To test for this the mode of each target and distractor piece was coded. Analysis of variance with Target type (Major/Minor) crossed with Discrimination type (Same/Different mode) again showed no evidence of a main effect on proportion correct of either target type ($F_{(1,11)} = 0.32$, p=0.58) or discrimination type ($F_{(1,11)} = 2.87$, p=0.12) and no interaction. The analysis for average response times also provided no evidence of an effect of either target type ($F_{(1,11)}=1.312$, p=0.28) or discrimination type ($F_{(1,11)}=0.87$,p=0.37) and no interaction.

Drawing Types The drawings produced fell into two broad categories. The first category of drawings, 'Abstract' (illustrated in Figure 1), involve some representation of musical form, e.g., intensity, pitch, melody, rhythm or tempo, typically represented as a contour. Attempts to use formal music notation were also coded as Abstract. It was notable, however, that use of formal notation was rare and, except in once case, subjects did not

[1] An alpha level of .05 was used for all statistical tests.

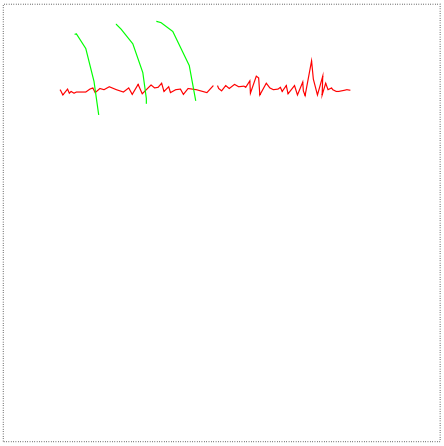


Figure 1: Example Abstract Drawing

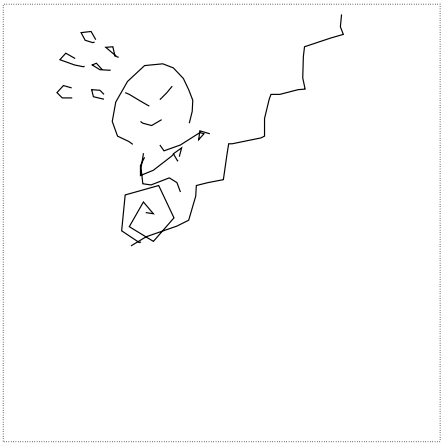


Figure 2: Example Figurative Drawing

persist with it. This was probably attributable to the difficulty of real-time transcription and the fact that it is only useful if both members of a pair are sufficiently expert with it. The second category of drawings, 'Figurative' (illustrated in Figure 2), is a more heterogeneous category involving depictions of e.g., faces, figures, objects or situations. A third, smaller category of Composite drawings was noted in which some abstract and figurative elements had been combined.

Two of the authors independently coded 287 drawings as either Abstract, Figurative or Composite. Inter-judge agreement on the coding was high (Kappa = 0.9, N = 287, k= 2).

It was initially hypothesised that the use of Abstract or Figurative drawings types might be influenced by the manipulation of genre. Jazz is often considered a more abstract form than Classical music and this might be reflected in the use of Abstract representations. Conversely, the selection of non-diatonic Jazz pieces containing some improvisation ensured that, relative to the Classical pieces, they had a less regular structure. On these grounds the form of Classical targets might be more easily detected and drawn. However, as Table 1 indicates, there appears to be no pattern in the distribution of drawing types according to the genre of the pieces. An analysis of the frequency of Abstract and Figurative drawings for Jazz and Classical pieces respectively suggests no reliable difference in pattern of use ($Chi_{(1)}$= 0.45, p=0.50).

Table 1: Frequency of Drawing Types According to Musical Form

	Drawing Type		
	Abstract	Figurative	Composite
Classical	37	94	12
Jazz	42	89	12

To test for effects of communicative context on choice of drawing type, measures of entrainment or matching were used (see Garrod & Anderson, 1987). This is the number of drawings produced by an individual that are of the same type (in this case Abstract, Figurative or Composite) as the immediately preceding drawing produced by their partner. It indexes the degree to which the members of a pair are tending to coordinate their choice of drawing type over and above what would be expected by chance given the frequency of use across the population as a whole. The average score for pairs in experiment 1 was 0.71 compared with a chance level of 0.49 (chance is calculated as the sum of the squared proportions of each drawing type in the corpus as a whole). These were reliably different ($t_{(11)}$=4.03, p (2 tailed) = 0.00).

Discussion of Experiment 1

Although the results show that subjects are able to carry out the task they provide no evidence of an effect of domain structure on task performance. Neither musical genre nor mode, as operationalised here, influenced the effectiveness with which pairs could perform the communication task. The difficulty of producing a drawing, as assessed by the effects of target type, was unaffected by musical form. The difficulty of distinguishing between pieces, as assessed by the effects of discrimination type, was also unaffected. Of course, it is possible that the experiments were insufficiently sensitive to detect an effect or that other aspects of musical form be influencing performance. Nonetheless, two intuitively salient aspects of form; genre and mode, did not affect the difficulty of the task.

More importantly, although the drawings can be reliably classified as Abstract or Figurative, the distribution of these representation types shows no influence of musical form, Jazz and Classical pieces are equally likely to be drawn in a Figurative or Abstract style. The form of the drawings is, however, predicted by communicative context. Pairs tend to entrain to one another, producing

matching drawing types more frequently than would occur by chance. These findings are consistent with our proposal that communicative coordination provides one of the principal influences on choice of representation type. However, the level of communication possible in experiment 1 was very limited. The interaction consisted only in the alternation between the roles of Giver and Drawer and the feedback about whether the last drawing had been correctly identified at the end of each trial. In effect, each trial is analogous to a single conversational turn. If communicative use constrains representational form then manipulation of the level of interaction should affect choice of representation type.

Experiment 2 was designed to address two issues. Firstly to investigate whether altering the richness of the communicative exchange would affect use of drawing types. Secondly to investigate the prediction that a medium which tended to discourage fluent production of graphics would favour more reduced, abstract, forms.

Experiment 2

In experiment 2 the richness of the communicative interaction was increased by allowing both participants to draw and erase freely on the shared whiteboard at any time. The manipulation of medium was introduced by contrasting two conditions, one with the same stylus based interaction as experiment 1, in which subjects draw directly onto the screen, and one with mouse based input. This served to reduce fluency of movement and introduced a spatial separation between input and the screen.

Materials A total of 112, 30 second, piano solo pieces, were used. This included the pieces used in experiment 1 as a subset. The pieces were selected according to the same criteria as in experiment 1 with the exception that each composer was used twice.

Subjects 24 participants, (16 male and 8 female, average age 19) were recruited from a variety of disciplines at local colleges and universities. They were paid an honorarium for taking part.

Procedure Broadly the same procedure as experiment 1 was followed. However, because the restriction to a single person drawing on the whiteboard was removed, subjects were shown a demonstration of simultaneous drawing on the whiteboard. Two versions of the interactive experimental task were used. In the 'matching' version both members of a pair had one piece of piano music each and the task was for them to determine whether these pieces were the same or different. In the 'discrimination' version each member of the pair had two pieces of music and the task was for them to decide which of the two pieces was the same. Although we do not discuss the task manipulation here, the analysis reported below is based on data from both tasks to preserve the balance of conditions and materials.

Design Experiment 2 employed a within-subjects, factorial design with task (Matching vs. Discrimination) crossed with Media (Mouse vs. Stylus). Selection of music was constrained so that the combinations of form (Jazz, Classical) and Mode (Major vs. Minor) were counterbalanced across conditions. Each piece was also classified according to its tempo with selection of tempo randomised across conditions. As before, no one heard the same piece of music twice. This design resulted in a total of 68 trials per pair with order of conditions and materials counterbalanced.

Results

The effects of media and experience on task performance were assessed in two analyses of variance. To index experience, the trials were divided into four blocks. For each quarter of the experiment, and each pair, the proportion of correct responses and the average time to respond were calculated. Analysis of variance on the proportion of correct responses, with media and experience as within subjects factors, showed a reliable main effect of experience ($F_{(3,24)} = 4.84$, p=0.01) but no effect of media ($F_{(1,24)} = 0.04$, p=0.84) and no interactions. Linear trend analysis confirmed that participants became more accurate with experience ($t_{(33)} = 3.48$, p(one tailed) = 0.00) . The parallel analysis for time to respond also showed a main effect of experience ($F_{(3,47)} = 4.07$, p=0.18) and again, no effect of media ($F_{(1,47)} = 0.01$, p=0.92). Linear trend analysis confirmed that participants were becoming faster at the task with experience ($t_{(33)} = 2.38$, p(one tailed) = 0.01). The results suggest that the manipulation of medium does not affect participants' ability to carry out the task.

Drawing Types The drawing activity of each member of a pair was separated into two files and independently coded, as before, for the categories Abstract, Figurative or Composite. An additional category of 'None' was introduced to deal with a small number of cases (3%) where one or both of the partners had not drawn a picture of a piece on a given trial. The distribution of each drawing type across all trials is given in Table 2.

The prediction that medium should affect distribution of drawing types was assessed by scoring, for each pair, the proportion of drawings that were classified as Abstract. This was analysed in an analysis of variance with medium, task and experience as within subjects factors. There was no simple main effect of media ($F_{(1,24)} = 2.22$, p= 0.15) and no reliable interactions.

Table 2: Distribution of Drawings, Experiment 2.

	Drawing Type		
	Abstract	Figurative	Composite
Frequency	970	345	257
Proportion	59%	21%	16%

Entrainment scores were calculated, as before, to provide an indication of the extent to which the members of a pair were coordinating their choice of representation

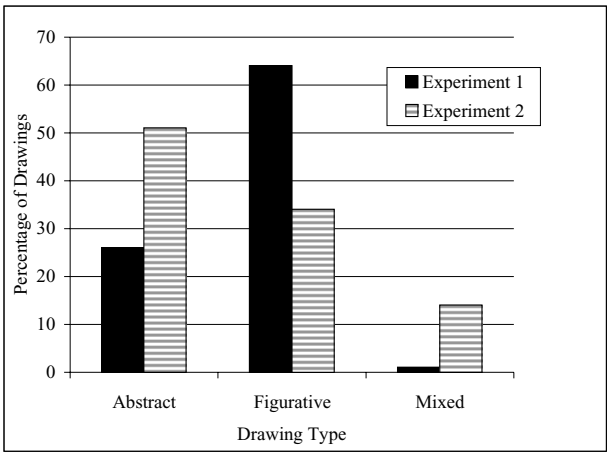


Figure 3: Choice of Drawing Types in Experiments 1 and 2

type. In this case scores were calculated as the proportion of trials in which drawings of the same type were produced, excluding trials in which one or both participants produced no drawing. The average entrainment score was 0.79, reliably above the chance level of 0.42 ($t_{(11)} = 6.75$, $p = 0.00$).

Comparison of Experiments 1 and 2

To test for effects of the difference in level of interactivity between experiments 1 and 2 only data from the first 12 trials of experiment 2 were used. This was in order to restrict comparisons to the situation in which participants had completed the same number of drawings of different pieces.

To compare level of coordination in choice of drawing type a t-test was performed on the average entrainment scores for each pair with experiment (1 vs. 2) as a between subjects factor. This indicated that levels of matching were not reliably different ($t_{(22)} = 0.74$, p (2 tailed) = 0.46). The ability to interact directly did not affect the extent to which pairs tended to match their choice of representation type.

Although degree of matching did not differ between experiments 1 and 2, a reversal in patterns of choice in drawing types was observed. As Figure 3 illustrates, during the first 12 trials of experiment 2 almost twice as many Abstract drawings were produced than in experiment 1. The contrast in relative frequency of Abstract and Figurative drawings confirmed the reliability of this pattern ($Chi^2_{(1)} = 50.7$, $p = 0.00$). The results indicate that the ability to interact directly has a substantial effect on the use of drawing types leading, in particular, to a much greater use of Abstract drawings.

General Discussion

Considered together, the results provide evidence that communicative use has a strong effect on representational form. Although intuitively genre and mode are important elements musical form, they had no effect on task performance or on choice of representation type in the present study. Additionally, although the manipulation of medium between mouse-based and stylus based input provides a contrast in levels of control and fluency it had no demonstrable effect on either performance or representational form. In particular, no evidence was found for the prediction that the simpler contours of Abstract drawings would be favoured when subjects used a mouse. In contrast to medium and domain, two effects of communicative use were noted. Firstly, subjects' choice between Abstract and Figurative representations was sensitive to their partner's choice of representation. People were much more likely to produce a drawing of a similar type to the one last produced by their partner than could be expected by chance. This pattern of entrainment parallels findings for dialogue. Garrod and Anderson (1987) found that, while domain structure favours some types of verbal description over others, the main constraint on choice of representation type is the pressure to coordinate with an interlocutor. Secondly, the pattern of use of drawing types between experiment 1 and 2 effectively reverses with approximately twice as many Abstract drawings and half as many Figurative drawings used in experiment 2. This suggests that level of interaction has an especially marked effect on choice of drawing type and this occurs even though the level of entrainment or matching in the two experiments is not reliably different.

One potential issue with this interpretation is that additional pieces of music were used in the second experiment, raising the possibility that these pieces particularly favoured Abstract drawings. However, the selection criteria for pieces were the same across both experiments and it seems unlikely that a specific bias was introduced. A second issue is that the difference in level of interaction is not the only difference between the tasks in experiments 1 and 2. Although number of pieces drawn and taken into account in the analysis, other task differences might have contributed to the observed effect on drawing types. Arguably, the interactive versions of the task are more comparative because more than one piece is drawn on each trial. This would not explain the tendency to entrain but it could contribute to choice of drawing type. This possibility is being investigated in further work.

The issues raised above notwithstanding, the results provide evidence of a substantial influence of communicative constraints on representational form. This does not necessarily undermine the claim that domain structure and media type influence representational form. Amongst other things, it is possible that genre and mode are relatively unimportant aspects of musical form and other aspects of domain structure would have a more marked effect. Similarly, the difference in ease of execution between a mouse and a stylus, although significant, may be insufficient to affect representational form. These questions can only be resolved by further empirical work. However, the present study suggests that constraints deriving from communicative use can have a strong, per-

haps key, influence on representational form.

The interpretation of these results depends on providing an account of what the difference between Abstract and Figurative drawings consist in. The Abstract category consisted of drawings that appeared to pick up on formal aspects of the music. Contour lines and blobs were used to represent a potentially wide variety of possible regularities, e.g., pitch, stress, harmonic structure, chord structure, rhythm, tempo, texture and intensity. For each of these possibilities a further number of variations are possible including choice of axes, choice of scale, and level of granularity (whole piece, first few bars etc.). A specific type of Abstract drawing imposes a systematic interpretation. It generalises to any piece of music and can sustain internal structural inferences for a piece, e.g., one chord is twice as long or intense as another. Relative to Abstract drawings, Figurative drawings, are highly heterogeneous. They employ a range of ad hoc devices such as visual emblems (city skylines for Jazz) symbols of emotive affect (sad faces, graves), pictures of rabbits or cars to indicate tempo, pictures of landscapes to suggest moods and so on. In contrast to Abstract drawings they provide only weak support for generalisations. There is not a street scene or landscape for every piece and they provide almost no information about the internal structure of a piece.

The present proposal is that the key constraint on the use of Abstract or Figurative drawings in the present task is the degree of coordination they require. In particular, successful use of Abstract drawings demands a higher degree of semantic coordination. To use Abstract drawings successfully subjects must attempt to coordinate on which aspect of musical form is being used, on which axes, at what level of granularity. Figurative drawings, by contrast, can be used in a more ad hoc manner. They can exploit different interpretations in each case and do not impose a particular structure on the music. We propose that direct interaction sustains the use of Abstract drawings by providing mechanisms that facilitate the coordination of interpretation. For example, in experiment 2 subjects were seen to circle and underline parts of each other's representations. This could function as a means of isolating and repairing problems with particular elements of a drawing. Lines and arrows between different contours were also used to indicate possible alignments or changes of scale. In experiment 1 such exchanges were impossible even though, in principle, the same types of representation could have been used.

The implication of these considerations is that properties of representations, such as abstraction, may depend more on the character of the interactions in which they are used than on the character of the represented domain. This is not to suggest that effective representations do not address regularities in the represented domain but rather that representational form is conditioned, first and foremost, by the structure of interaction and the kinds of coordination that it makes possible, and only contingently by the structure of the domain.

Acknowledgments

We wish to thank ATR Media Integration and Communications Laboratories, Kyoto Japan and the ESRC and EPSRC (grant L328253003) for generous support. A previous version of this paper was presented under the title: "Graphical Interaction and the Emergence of Abstraction" at the First International Workshop on Interactive Graphical Communication, Queen Mary, University of London. Aug 30-31st, 2000.

References

Bartlett, F. (1932). Remembering: A study in experimental and social psychology. *CUP*.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*, 181-218.

Garrod, S., & Doherty, G. (1994). Conversation, coordination and convention: An emprirical investigation of how groups establish linguistic conventions. *Cognition*, *53*, 181-215.

Healey, P. G. (1997). Expertise or expert-ese: The emergence of task-oriented sub-languages. In M. Shafto & P. Langley (Eds.), *Proceedings of the ninteenth annual conference of the cognitive science society* (pp. 301–306). Stanford University, CA.

Healey, P. G. (2001). *Semantic coordination in dialogue.* (Manuscript in preparation)

Scaife, M., & Rogers, Y. (1996). External cognition: How do graphical representations work? *International Journal of Human-Computer Studies*, *45*, 185-213.

Schwartz, D. L. (1995). The emergence of abstract representations in dyad problem solving. *The Journal of the Learning Sciences*, *4*(3), 321–354.

Shimojima, A. (1996). Operational constraints in diagrammatic reasoning. In G. Allwein & J. Barwise (Eds.), *Logical reasoning with diagrams* (pp. 27–48). Oxford.

Stenning, K., & Oberlander, J. (1995). A cognitive theory of graphical and linguistic reasoning: Logic and implementation. *Cognitive Science*, *19*, 97–140.

Tversky, B. (1981). Distortions in memory for maps. *Cognitive Psychology*, *13*, 407–433.

Tversky, B. (1989). Perceptual and conceptual factors in distortions in memory for graphs and maps. *Journal of Experimental Psychology: General*, *118*(4), 387–398.

Tversky, B. (1995). Cognitive origins of graphic conventions. In F. Marchese (Ed.), *Understanding images* (pp. 29–53). New York: Springer-Verlag.

Pragmatics at Work: Formulation and Interpretation of Conditional Instructions

Denis J. Hilton (hilton@univ-tlse2.fr)
Laboratoire D.S.V.P., Université Toulouse 2
5 allées A. Machado, 31058 Toulouse cedex, France

Jean-François Bonnefon (bonnefon@univ-tlse2.fr)
Laboratoire D.S.V.P., Université Toulouse 2
5 allées A. Machado, 31058 Toulouse cedex, France

Markus Kemmelmeier (markusk@umich.edu)
Department of Psychology, University of Michigan
525 E. University Ave., Ann Arbor, Michigan, 48109-1109, USA

Abstract

Formulation and interpretation of conditional instructions (conditionals relating the occurrence of an event to the taking of an action) are studied from a pragmatic standpoint: It is argued that formulations of the instructions differ in perceived naturalness as a function of the adequacy between the necessity and sufficiency relations they embed and the goal-structure of the situation. Two experiments are reported to support this claim.

Conditional Instructions as a Peculiar Subclass of Conditional Statements

As Austin (1962) has observed, we use words to get things done – and there are indeed lots of things we can get done by using the word "if", that is, by asserting a conditional statement. With a conditional statement, we can tell others what conclusion they should draw (e.g., "if it is a three-star restaurant, then the food there is certainly divine"), or what action they should take (e.g., "if you wear this suit, you will make a very good impression on her") – and in particular, especially in working situations, we can give others instructions to follow would a specific situation occur (e.g., "if a customer buys two make-up products, offer her a sample of this perfume").

It is common knowledge among psychologists that people do not do very well in conditional reasoning tasks. For example, they have a disturbing tendency to derive the wrong conclusions from a conditional argument (e.g., they commit the fallacies of Asserting the Consequent: "If P then Q; Q; therefore P," and of Denying the Antecedent: "If P then Q; not-P; therefore not-Q") and not to derive the right ones (e.g., they do not apply Modus Tollens: "If P then Q; not-Q; therefore not-P"). (See Evans, Newstead, & Byrne, 1993, for a review.) Does this poor performance extend to this subclass of conditional statements we just dubbed conditional instructions? Since conditional instructions are so very common in working situations, that could be bad news.

Fortunately, people seem to be better at handling conditional instructions than at solving conditional reasoning problems. In this paper, we want to show that when dealing with conditional instructions, people are able to (a) select the better way to express a conditional relation between two events according to the goal-structure of the context, and (b) interpret a conditional relation in a normatively (logically) valid way.

Goal-Structure of the Context and Formulation of a Conditional Instruction

From our pragmatic point of view, the context will determine what a speaker will aim to achieve by uttering a conditional instruction: Thus, conditional instructions make points, and do so well or badly depending on their perceived relevance to contextually specified goals. Context will be considered here through an analogy between complying with a conditional instruction and being engaged in a signal detection task. (See Kirby, 1994, for another view on the analogy between conditional reasoning tasks and signal detection tasks.)

In a signal detection task, the observer has to judge whether a signal actually indicates its putative referent or not. Does a certain kind of blip on a sonar screen indicate the presence of an enemy submarine (that has to be sunk) or not? Two types of error are conceivable here: (a) other kinds of objects may have caused the blip, for example whales or friendly submarines, leading to a "false alarm" (FA); and (b) an enemy submarine may really be out there, his sonar "signature" being distorted by underground wave patterns or rock formations in such a way that the operator fails to recognize it, leading to a "miss" (MS).

Now, whether a MS or a FA has the highest expected cost may depend on the situation. In a state of war, MS are likely to prove costly: If you do not sink the enemy submarine first then it will sink you. However, a FA may also prove costly: If you mistakenly sink a neutral country's submarine or ship, you may provoke that country to declare war on you, which could prove especially costly if the neutral country happens to be, say, the United States. (Think of the Lusitania...)

Now, imagine a warship commander wishing to give his operators a conditional instruction linking the observing of an enemy submarine "blip" and the launching of depth charges. In context A, the commander knows that enemy submarines are lethal if allowed within range and must be destroyed at first sighting. Clearly, what he should fear are MS, that is, enemy submarines which are not attacked. In context B, the commander knows that enemy submarines are outside range at first sighting, and that there is a considerable risk of destroying his own submarines, or those of a neutral superpower that are also lurking in the area. What he should fear here are FA, that is, non-enemy submarines which are attacked.

What would be the best way for the commander to frame his conditional instruction in context A? "If you see an enemy blip, then launch the depth charges?" "If and only if you see an enemy blip, then launch the depth charges?" "If you do not see an enemy blip, then do not launch the depth charges?" "Launch he depth charges only if you see an enemy blip?" And what would be the best choice in context B?

Our prediction is that people can and do perceive differences in the naturalness of these formulations as a function of what they perceive to be the goal-structure of the context, that is, "avoid misses" vs. "avoid false alarms". Experiment 1 below offers an experimental investigation of this claim.

Experiment 1

Participants A total of 46 students at the University of Heidelberg took part in this study.

Material & Method Three additional scenarios were created on the model of the Submarine scenario: each scenario came either in an avoid-FA context or in an avoid-MS context. In the Airport scenario, a security officer was to decide whether he would search suspicious-looking luggage, knowing that (avoid-MS context) the airport was situated within a "hot" area where terrorists were liable to smuggle weapons, or (avoid-FA context) the airport was mostly frequented by high-ranking executives that would not appreciate losing their time with a luggage search. In the Border scenario, a policeman equipped with a speed radar of some poor quality had to decide whether he would arrest drivers slightly exceeding the speed limit when entering France from Germany via a subway (the speed limit in France being lower than in Germany), knowing that (avoid-MS context) officials insisted on strictly implementing the French regulation, or (avoid-FA context) the officials insisted on the importance of fluent circulation prior to the strict implementation of the French regulation. In the Mail scenario, which was adapted from Gigerenzer and Hug (1992) in order to vary costs and benefits from a single perspective, an office worker was told to stamp letters over 20 grams in weight at 2 marks, knowing that (avoid-MS context) understamping (i.e., putting 1 mark stamps on letters over 20 grams) would damage the firm's public image, or (avoid-FA context) overstamping (i.e., putting 2 marks stamps on letters under 20 grams) would be costly to the firm's finances.

Each questionnaire featured the four scenarios, all of them in their avoid-MS or avoid-FA version (context was thus a 2-level between-subject factor), rotated over two experimental blocks, with two orders within each experimental block, such that each subject saw two avoid-MS and two avoid-FA contexts paired with different content scenarios. Following each scenario, four conditional instructions were introduced (if P then Q, if and only if P then Q, Q only if P, if not-P then not Q); participants had to rate on a 7-point scale the naturalness of each instruction in the situation that had just been described to them (formulation of the conditional instruction was thus a 4-level within-subject factor). The experiment was conducted in German.

Results & Discussion Table 1 displays the mean naturalness ratings (across the four scenarios) assigned to the four formulations as a function of the goal-structure of the context. (The observed pattern of results was remarkably stable across scenarios.)

Table 1: Naturalness ratings (7-point scale) of conditional formulations as a function of context.

	Context: avoid-MS	Context: avoid-FA
If P then Q	5.74 [a]	3.10 [c]
If and only if P then Q	4.03 [b]	4.47 [b]
Q only if P	3.19 [c]	4.33 [b]
If not-P then not-Q	2.42 [d]	4.91 [b]

N = 22 for avoid-MS context, N = 24 for avoid-FA context. Values that do not share the same subscript differ at $p < .05$.

In the avoid-FA context, all formulations appear to be of acceptable naturalness (4 to 5 on a 7-point scale), except the "if P then Q" formulation which is judged significantly less felicitous. On the contrary, this formulation is by far the most felicitous in the avoid-MS context, the formulations "Q only if P" and "if not-P then not-Q" being this time judged unnatural.

Now why these differences in naturalness as a function of context? One possible answer is related to the notions of necessity and sufficiency. In the avoid-MS context, one would like to stress the sufficiency of P (observing an enemy blip) in regard to Q (launching the depth charges), whereas in the avoid-FA context, stress should be on the necessity of P in regard to Q. Hence, the ideal formulation in the avoid-MS context would be "if P then Q", whereas this same formulation would be inappropriate in the avoid-FA context. In a given context, a natural formulation will be one that direct the attention of the hearer to the relevant aspects of the situation: is P necessary rather than sufficient for Q? This explanation assumes that people's interpretation of the necessary and/or sufficient character of P in regard to Q in the four considered formulations coincide with what it should be according to formal logic. In the light of previous research (see again Evans, Newstead, & Byrne, 1993), this could be seen as a rather bold assumption. The next section will focus on the reasons why this assumption may hold in the specific case of conditional instructions.

Interpretation of Conditional Instructions as Constraint Perception

Does each of our four conditional formulations (if P then Q, if and only if P, Q only if P, if not-P then not Q) have its own stable interpretation in terms of necessity and sufficiency relations? That is, do people consider these formulations to embed different basic patterns of necessity and sufficiency, even if they have no idea of the goal-structure of the instruction? Moreover, do these patterns coincide with the patterns predicted by traditional logic?

The standard approach to this issue would have been to give participants a scenario (e.g., selling clothes in a clothing store), an instruction (e.g., "if a customer is not touching any clothes, do not offer him your help"), a situation (e.g., "a customer is touching some clothes"), and ask them what they would do in this situation if they had to follow the rule (e.g., "I would offer my help", "I would not offer my help", "I do not know"). But this approach would actually miss the point, for it would not assess the interpretation subjects made of the rule, but their final decision on what they should do, a decision that does not solely depend on the interpretation they made of the rule. (In the above example, a participant may well answer that she would offer her help to a customer that is touching some clothes, after being told that "if a customer is not touching any clothes, do not offer him your help". Is this participant interpreting the rule as meaning that a customer touching some clothes is a sufficient condition to offer him some help? Or is she just taking her best bet on what to do when the rule does not strictly apply?)

Therefore, in order to assess the interpretation participants make of a conditional instruction, what has to be checked is not what they would do in the situations P and not-P, but how they perceive the way the instruction is constraining their behavioral options in these situations. Thus, given the rule "if a customer is not touching any clothes, do not offer him your help", and the situation "a customer is touching some clothes", the relevant set of answers to choose from would be: "I must offer my help", "I must not offer my help", and "I am free to decide what to do."

We proposed that the function of the different formulations of the instruction was to direct attention

Table 2: Most frequent patterns associated to each formulation of the conditional instruction (N = 39).

Formulation	Most frequent pattern	Frequency: Shop scenario	Frequency: Restaurant scenario
If P then Q	Situation P: Must do Q Situation not-P: Free to decide	82%	82%
If and only if P then Q	Situation P: Must do Q Situation not-P: Must not do Q	85%	82%
If not-P then not-Q	Situation P: Free to decide Situation not-P: Must not do Q	85%	90%
Q only if P	Situation P: Free to decide Situation not-P: Must not do Q	56%	46%
	Situation P: Must do Q Situation not-P: Must not do Q	31%	46%

on different aspects of the context. Efficient illocutionary uptake would then depend on the possibility for the hearers to rely on some basic, conventional meaning of the four formulations regarding the necessity and sufficiency relations they embed. Were these basic meanings to coincide with what they are in traditional logic, then given a conditional instruction "if P then Q", "if and only if P then Q", "Q only if P", or "if not-P then not Q", and the set of choices "I must do Q", 'I must not do Q", "I am free to decide what to do", participants' answers in the situations P and not-P would exhibit normative (logical) validity in terms of the necessary and/or sufficient relationships between P and Q. Experiment 2 below was designed to provide an empirical investigation of this hypothesis.

Experiment 2

Participants A total of 39 students of the Ecole Supérieure des Sciences Economiques et Commerciales (ESSEC) at Cergy-Pontoise took part in this study.

Material & Method Two scenarios were constructed, the Shop scenario and the Restaurant scenario. In the Shop scenario participants were told that they were selling clothes in a shop; they had to decide whether they would offer a customer some help, knowing that there was an instruction to be strictly followed (e.g., "if a customer is touching some clothes, offer him some help"). In the restaurant scenario, participants were told they were establishing a list of providers for the chef; they had to decide whether a provider should be put on the list, again knowing that there was an instruction to be strictly followed (e.g., "if a provider does not offer you a reduced price, do not put him on the list").

Each questionnaire featured the Shop scenario and the Restaurant scenario. Within each scenario, the four formulations of the conditional instruction were introduced in turn. (For the Shop scenario, the four formulations went:
"If a customer is touching some clothes, offer him your help", "If and only if a customer is touching some clothes, offer him our help", "Offer a customer your help only if he is touching some clothes", and "If a customer is not touching any clothes, do not offer him your help.") The formulation of the instruction was thus a 4-level within-subject factor. For each rule, participants were asked to choose from the three following answers, first in the situation P, then in the situation not-P: "I must do Q", "I must not do Q", "I am free to decide what to do." The experiment was conducted in French.

Results & Discussion A first way to look at the results is to consider the most frequent pattern of answer elicited by the participants for each formulation (see Table 2). Regarding the formulations "if P then Q", "if and only if P then Q", and "if not-P then not-Q", there is a clear dominance of a single pattern for each rule (eliciting 82 to 90% of answers), whereas the formulation "Q only if P" elicits two main patterns. (Whatever the formulation, no other pattern elicited more than 13% of answers.) The dominant patterns elicited by the formulations "if P then Q", "if and only if P then Q", and "if not-P then not-Q" are precisely those that would be predicted by classical logic. Of the two main patterns elicited by the formulation "Q only if P", one is predicted by classical logic, the other one is the biconditional pattern.

Another way to look at the results is to consider, for each formulation of the instruction, the frequency with which participants answered as if P was necessary (see Table 3) or sufficient (see Table 4) for Q. In order to compute the percentages in Tables 3 and 4, participants have been considered as (a) answering as if P was necessary for Q if they answered that they would have to avoid doing Q in the situation not-P, and (b) answering as if P was sufficient for Q if they answered that they would have to do Q in the situation P.

Whatever the scenario, P was overwhelmingly considered to be necessary for Q with all formulations except "if P then Q", which is what one would expect according to classical conditional logic. In particular, the fallacy of Denying the Antecedent ("if P then Q, not-P, therefore not-Q") was endorsed by only 8 to 13% of the participants, which is well below the usual rate observed in conditional reasoning experiments.

Table 3: Necessity of P in regard to Q (in percentage of answers), as a function of instruction formulation.

	Shop scenario	Restaurant scenario
If P then Q	08 % [a]	13 % [a]
If and only if P then Q	95 % [b]	95 % [b]
Q only if P	87 % [b]	92 % [b]
If not-P then not-Q	92 % [b]	97 % [b]

N = 39. Values that do not share the same subscript differ at $p < .05$.

Turning to the sufficiency of P in regard to Q, results are unambiguous for the formulations "if P then Q", "if and only if P then Q", and "if not-P then not-Q": This time, P is overwhelmingly deemed as sufficient for Q, as classical logic would predict. The unexpected result (from a logical standpoint) comes from the formulation "Q only if P", with P being deemed sufficient for Q by 36 to 51% of participants. This last result may be used to rule out the idea that logical competence only could be responsible of participants' answers: If participants recovered logical competence when dealing the instructional subclass of conditional statements, then

why would the specific formulation "Q only if P" elicit logical errors?

Table 4: Sufficiency of P in regard to Q (in percentage of answers), as a function of instruction formulation.

	Shop scenario	Restaurant scenario
If P then Q	87 % [a]	92 % [a]
If and only if P then Q	87 % [a]	85 % [a]
Q only if P	36 % [b]	51 % [b]
If not-P then not-Q	08 % [c]	08 % [c]

N = 39. Values that do not share the same subscript differ at $p < .05$.

Without resorting to an explanation in terms of logical competence, it could be argued that the deontic nature of conditional instructions is responsible for the normatively correct performance of participants, since deontic contents are known to be a powerful facilitator of conditional reasoning. First, it should be noted that a conditional instruction is not a social contract the way Cosmides (1989) has defined it: A conditional instruction does not relate perceived benefits to perceived costs, it does not express a social exchange in which an individual is required to pay a cost (or meet a requirement) to another individual in order to be eligible to receive a benefit from that individual. Having no cost-benefit structure, conditional instructions do not leave room for cheating, that is, obtaining the benefit without paying the cost. Therefore, if participants' performance has benefited from some deontic facilitation, this facilitation does not fall within the scope of Cosmides' (1989) social contract theory or Gigerenzer and Hug's (1992) cheater-detection algorithm.

Would this deontic facilitation be explainable by Cheng and Holyoak's (1985) pragmatic reasoning schemas theory? In Cheng and Holyoak's terms, improved performance would be due to some content or context-based prompting of either a permission or an obligation schema. Yet, since context and semantic content of the instruction stay the same across our conditions, why should syntax alone determine the nature of the prompted schema? We fail to see why, content and context remaining stable, "if P then Q" would lead to the activation of an obligation schema, whereas "if not-P then not-Q" or "Q only if P" would lead to the activation of a permission schema.

As demonstrated by Thompson (2000) in her study of interpretative processes in various types of conditional reasoning tasks, performance in a conditional argument task (contrary to performance in Wason's selection task) is predicted by necessity and sufficiency conditions, and not by the deontic or factual nature of the conditional. Is it possible to explain our results in terms of perceived necessity and sufficiency relations?

Indeed, conditional instructions are meant to embed very strong necessity and sufficiency relations: In the instruction "put a provider on the list only if he offers you a reduced price", the necessity of the offer is clearly not a matter of degree. Due to the intrinsic nature of conditional instructions, any necessity or sufficiency relation between the two propositions involved will be of maximal perceived strength, which would explain the extreme frequencies observed in Tables 3 and 4. The fact that participants did so well in perceiving the valid necessity and sufficiency relations and dismissing the invalid ones in the instructions they were given can conceivably be explained by one distinctive aspect of conditional instructions: Contrary to most conditionals (e.g., causal conditionals, conditional warnings, etc.) instructions are not meant to change the epistemic state of their recipient, but to constrain his or her behavior. As one's natural preference will usually be to exert one's free will, it is not much surprising that one will be accurate in recognizing in which situation one's behavior will be dictated or not by the instruction, that is, recognizing the necessity and sufficiency relations embedded in the instruction.

Conclusion

The focus of this paper has been the formulation and interpretation of conditional instructions, that is, conditionals that relate the occurrence of some event to the undertaking of some action. Drawing an analogy from signal detection theory, we labeled a "Miss" the situation in which the event is occurring but the action is not taken, and a "False Alarm" the situation in which the action is taken without the event occurring.

We proposed that context allows to determine the relative expected costs of Misses and False Alarms, which in turn allows to determine the goal-structure of the situation and the aim of the speaker asserting the instruction: What is to be avoided in the situation? Misses? False Alarms? Both?

Depending of the goal-structure of the situation (and consequently of the aim of the speaker), syntactic formulations of the instruction differ in perceived naturalness (Experiment 1). For example, the usual conditional formulation "if P then Q" will be perfectly appropriate for situations where Misses must be avoided, but will be of poor felicitousness in situations where False Alarms must be avoided.

We proposed that judgements of naturalness are based on the understanding people have of the necessity and sufficiency relations embedded in the various possible formulations of the instruction. A formulation in which P is sufficient for Q is appropriate for situations where Misses must be avoided, a formulation in which P is necessary for Q is appropriate for

situations where False Alarms must be avoided. Experiment 2 showed that when dealing with conditional instructions, people have a much clearer understanding of those relations than what could have been expected from their usual performance in conditional reasoning tasks.

Taken together, these two studies suggest that speakers' perception of the felicity of different kinds of conditional expressions is strongly determined by goal-structure (avoid miss vs. avoid false alarm), and that hearer's reactions to these conditionals is well aligned with this goal-structure, even if hearers have no explicit knowledge of these goals. The results therefore suggest that the function of these different formulations of conditional instructions is to direct the hearer's attention to aspects of his decision-making situation that the speaker considers important. That hearers so successfully detect the speaker's intentions not only suggests – in the language of Austin – high illocutionary uptake, but also successful coordination of action by the speaker and the hearer.

For the rational speaker to get what he wants done with words, he should therefore choose a form of the conditional that encodes the goal-structure implicit in the context, in the knowledge that the hearer should react in a way that will fulfill his intention. Rationality here is thus social and pragmatic, determined by the successful coordination of the speaker and the hearer to achieve shared organizational goals.

Acknowledgments

The authors would like to thank Fouad El-Ouardighi for his help in data collection.

References

Austin, J. L. (1962). How to do things with words. Oxford: Clarendon Press.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. Cognitive psychology, 17, 391-416.

Cosmides, L. (1989). The logic of social exchange: has natural selection shaped how we reason? Studies with the Wason selection task. Cognition, 31, 187-276.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). Human reasoning: The psychology of deduction. Hove: Lawrence Erlbaum Associates.

Gigerenzer, G., & Hug, K. (1992). Domain specific reasoning, social contracts, and perspective change. Cognition, 43, 127-171.

Kirby, K. N. (1994). Probabilities and utilities of fictional outcomes in Wason's four-card selection task. Cognition, 51, 1-28.

Thompson, V. A. (2000). The task-specific nature of domain-general reasoning. Cognition, 76, 209-268.

The Influence of Recall Feedback in Information Retrieval on User Satisfaction and User Behavior

Eduard Hoenkamp (hoenkamp@acm.org)
Henriette van Vugt (vanvugt@cogsci.kun.nl)
Nijmegen Institute for Cognition and Information; Montessorilaan 3
6525SW Nijmegen, the Netherlands

Abstract

The unprecedented scale-up of the World Wide Web, and the number of people relying on it for information, make it inevitable to reassess the validity of the traditional metrics for quality of information retrieval (IR). Of these, the most widely used metrics are recall and precision.
Users can judge the precision of an information retrieval system by inspecting the retrieved documents. They cannot judge recall, however, which would involve inspecting the whole collection, thus obviating the IR system, and impossible in the case of WWW. How then, can we ascertain whether recall is a valid metric for the quality of an IR system as perceived by the end-user? In a carefully controlled experiment we presented users with a simulated web search engine. Besides the search results, the engine could give a (spurious) recall estimate, presented as a pie chart. We manipulated this recall feedback, and whether the information need was fulfilled with respect to quantification type (the number of documents requested). It seems that fulfillment is a better predictor of user satisfaction and behavior than precision and recall as used to evaluate IR systems. The results reported here also suggest that whereas recall may be a valid metric for designers and evaluators of IR systems, it may lack validity as a metric for search quality as perceived by the end-user.

Introduction

Barely a decade ago, techniques for information retrieval were still in the able hands of librarians, in the case of print, and of data base managers in the case of electronically stored information. The explosive growth of the World Wide Web has changed this situation dramatically and irrevocably. Since then, people in all walks of life depend on automated 'librarians' as provided by search engines such as Google, AltaVista, Yahoo, and many others. Oviously, the end-user of such a system wants information that is relevant, and wants it returned within a reasonable time. The latter is a matter of efficiency, and that is where most of the research effort has gone. For example: how to increase bandwidth, how to index documents, how to encode multimedia. Not surprisingly, the aspect of efficiency is dominated by computer science, and solid metrics are known for these technical aspects.

Effectivenes, on the other hand, can only be gauged by the users of an IR system themselves. We claim that IR is a golden opportunity for cognitive science, with its roots in both psychology and computer science. For this, researchers can pursue two avenues: one is to exploit cognitive principles in modeling the user, the other by evaluating traditional metrics of IR concerning effectiveness through experimental design. The viability of the former approach we demonstrated elsewhere (Hoenkamp, Stegeman, & Schomaker, 1999; Hoenkamp & de Groot, 2000). In this paper we give an example of the latter.

From the the early days of the library sciences until today many metrics have been proposed to evaluate the quality of information retrieval systems (Swets, 1963; Cleverdon, Mills, & Keen, 1966). These metrics are to measure how satisfactory the material is that the system retrieves (the output), with respect to the user's information need presented as a query (the input). After several decades, *recall* (proportion of relevant documents actually retrieved) and *precision* (proportion of the retrieved documents that are relevant) have stabilized as *normative* measures for the quality and thus comparison of IR systems. The evaluation of these metrics has itself become a subject of study regarding both their reliability (Buckley & Voorhees, 2000) and their validity (Hersch, Turpin, Price, Chan, Kraemer, Sacherek, & Olson, 2000). Yet, however respectable and theoretically sound the metrics may be for comparing IR systems, it does not make them automatically apropriate to predict the satisfaction of the end-user with such a system. And given that IR systems are eventually built not for the evaluators but for the end-user, we wanted to investigate whether these metrics are also *valid* measures for quality from *the perspective of the user* conducting the search.

Users can only fare on the documents actually returned, and not on the uncounted documents never found. And as users can determine the relevant documents only among those returned, they can determine precision, but not recall. In addition, if users want to refine a search or provide feedback, again they can only do so on the basis of the documents returned. As precision is the only parameter the user can be aware of, it is the more likely parameter to determine the quality of a search as perceived by the user. So precision can be validated in principle, as one predictor of the user's satisfaction with an IR system. As the user cannot observe recall, there cannot be a corresponding validation for recall. This ends the symmetry between the two metrics that their definitions suggest. Any hope for exploring the relationship between recall and search quality, as perceived by the end-user, would require restoring that symmetry. This is exactly what we

set out to do. In a moment we will describe an experiment where we provided users with recall feedback, and measured the influence on their satisfaction with search results and search machine, and with their subsequent search behavior. Also, the usefulness of recall feedback is measured. It is important to understand that the recall feedback was represented by a slice on a pie chart. The size of the slice was manipulated, and had no relation whatsoever to actual recall.

It is useful to look first at our intuitions in order to appreciate what we learned through the experiment.

Intuitions

For the evaluator of IR systems, the intuitive trade-off between recall and precision is well-known: High recall can be achieved trivially by returning all documents, as this will include all relevant documents. Obviously, this goes at the expense of precision as many irrelevant documents are returned as well. Similarly, high precision can be achieved by stringent conditions on relevance, at the cost of missing potentially relevant documents. The end-user has also intuitions about recall (which we will capture below under hypotheses 2 and 7): When a search engine returns many relevant documents, the recall is perceived as high (especially when the precision is high). That is, the user thinks that the search engine retrieved a large proportion of the relevant documents. Consequently, the user is satisfied with such a search engine. Conversely, if very few documents, or none at all are returned the recall is seen as low, and the user is less satisfied with the search engine. Note, however, that the actual recall can be opposite to these intuitions. Especially when recall feedback violates these intuitions, this should influence the user's satisfaction with the search engine.

Focusing on the user's satisfaction with the search results, we intuit that it will depend on the degree to which the user's information need is met, and not on the mere number of returned relevant documents. For example: if the user wants to know whether the latest "Harry Potter" is out, just one document could meet this information need. If he wants to know which retailer on the web has the lowest price or the fastest delivery for the book, he needs as many sites as possible to choose from. Finally, if he needs the name of a bookseller nearby, a few documents may suffice to find one. Following Cooper (1968) we refer to these numbers as the *quantification type* of an information need, and call them quantification type 1, 2, and 3 (for one, all, or several documents). We expect the user to be most satisfied with the search result if the number of relevant documents returned matches the quantification type, at a high precision rate (this intuition leads to hypotheses 3 and 8). The satisfaction with the *search engine* we gather will depend on the type, the documents returned and, as the system is evaluated a whole, the *recall* (this leads to hypotheses 4 and 9).

These intuitions about the hypothetical relationship between the satisfaction and the compromise between recall and precision are visualized in figure 1. The figure shows the three quantification types. For example, for quantification type 3, the user would be unhappy with only one relevant document, satisfied with, say five documents, and again less satisfied when many more are returned as they will subsume more and more be irrelevant ones. The figure represents cases with, say, 200 relevant documents. The numbers on the x-axis are fictitious but are meant to indicate recall and precision. From left to right recall increases and precision decreases (recall and precision can easily be calculated, using the numbers in the figure). At the top of each curve the information need is fulfilled at the highest precision rate. The figure represents our prediction that no universally best compromise between recall and precision exists, as satisfaction will depend on the number of documents needed.

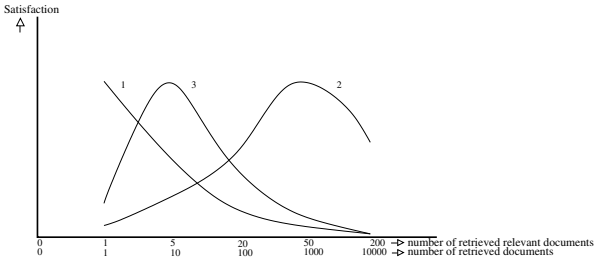


Figure 1: The compromise between recall and precision when *one* (1), *all* (2), or *several* (3) documents are needed. The curves represent the qualitative relationships hypothesized in this paper, between the satisfaction with retrieved documents and recall and precision. Note that the numbers along the x-axis are fictitious.

In this paper we assume that that user is looking for information, as opposed to entertainment. Hence we assume that the user's behavior, i.e. to continue or to stop searching, depends on whether his information need is fulfilled (which leads to hypotheses 5 and 10). Finally, we expect that also the usefulness of recall feedback depends on whether the the information need is fulfilled. More precisely, recall feedback will only be important if the users' need is not (yet) fulfilled. The usefulness may increase with increasing amount of documents searched for (these intuitions lead to hypotheses 6 and 11). we do not expect them to continue searching. Nor to

In line with the above, we expect strong interactions among the dependent variables. For example, when the information need of users is fulfilled, we expect them to be highly satisfied, to stop searching, and not to care much about recall feedback.

It will be clear that there are many potential interactions among the variables we introduced, hence the rather complex design. The hypotheses in the design section below are only a more detailed expression of the hypotheses we have just introduced informally.

Description of the experiment

Method

Participants The thirty two participants, fourteen female, were almost all acquaintances of one of the experimenters and volunteered to participate. Thirty of them had at least college education. They varied in age from 18 to 72, with a mean of 26. Their computer experience is presented in table 1.

Table 1: The participants' familiarity with computer activities. The columns indicate frequency the activity. a: never, b: once a year, c: once a month, d: once a week, e: several times a week, and f: daily.

	a	**b**	**c**	**d**	**e**	**f**
computer use	0	0	0	0	8	24
internet use	0	0	1	3	12	16
use of search engine	1	0	4	7	13	7

Design The experiment followed a within subjects design. Three variables were manipulated (1) fulfillment of the information need (2) quantification type (one, all, several), and (3) presence or absence of recall feedback, represented by a pie chart. If feedback was present, three ranges were used: low, medium, and high. These were depicted as slices of respectively 10, 20, and 30% of the pie, 40, 50, and 60%, and 70, 80, and 90%. Let us reiterate that the recall value had nothing to do with actual recall. It was only used to give the subjects the impression that the search engine produced this value. In reality a search engine cannot give such a precise number to a user, that would be a paradoxical situation where it would need the user to evaluate the relevance of documents it has not shown to the user.

In our pilot study we had prepared the material such that for each query we could give a number of relevant documents to match the quantification type. As the search engine would always return ten documents, we in effect controlled the *precision@10* (the proportion of relevant documents in the first ten documents).

The dependent variables were (1) satisfaction with the documents, (2) satisfaction with the search engine, (3) usefulness of the chart, and (4) tendency to continue to search. In a questionaire, the first three varialbels were scored on an 11-point Likert scale and the fourth was the answer to a yes/no question.

The hypotheses we investigated are an elaboration of the intuitions we described previously. Especially because they are so intuitively appealing, they have to be carefully laid out.

H1 Fulfillment of information need will be the dominating factor influencing the dependent variables.

Hence we split the other hypotheses up in two cases.

When the information need is fulfilled:

H2 The intuition of participants has the following effect: high recall causes a higher satisfaction, a higher usefulness of recall, and a higher stop rate than low or average recall.

H3 The satisfaction with the documents is high irrespective of quantification type and recall feedback.

H4 The satisfaction with the search engine is high and increases with magnitude of recall. There is no influence of quantification type.

H5 Users do not want to continue to search. Yet, a low recall may persuade them to do so.

H6 The usefulness of recall feedback is low and does not depend on its magnitude. If it would change at all, it would increase in the order of quantification type 1, 3, and 2.

When the information is not fulfilled:

H7 The intuition of participants has the following effect: low recall causes a lower satisfaction, a higher usefulness and a lower stop rate than average or high recall.

H8 The satisfaction with the documents is low irrespectively of quantification type and recall feedback.

H9 The satisfaction with the search engine is low, but increases with magnitude of recall. There is no influence of quantification type.

H10 Users want to continue to search. Yet, a high recall may persuade them to stop searching.

H11 The usefulness of recall feedback is high and does not depend on its magnitude. If it increases at all, it would be in the order of quantification type 1, 3, and 2.

Apparatus Participants interacted individually with Netscape 4.7 on a Macintosh G3. The HTML pages used in the experiment were stored locally to avoid network delays. Several toolbars ('navigation', 'location', and 'personal') were turned off to maximize window area as well as prevent interfering or unneeded interaction. The simulated search engine had the unadorned look and feel of the 'Google' search engine. The advantage of the simulation is obviously that all variables could be carefully controlled. Besides the query page, there was a page with search results (including documents and possible recall feedback) and a questionnaire existing of four questions and a box in which remarks could be written. For each search task we returned exactly ten documents. The participants were provided with pencil and paper to jot down the search task at hand. It had a circle printed on it, where they could copy the pie chart.

Procedure Each participant completed one practice task, and 24 randomized experimental search tasks that included a broad range of topics. The quantification type of each search task was obvious (e.g. the task to find a particular home page, is of type 1). The participants had to read the instructions from the screen. They were told

that we wanted to evaluate a search engine that used a novel search strategy. After the instructions, they had to explain the meaning of a pie chart, so we could check whether it was correctly understood (namely as recall information). For each task they went through the following cycle: (1) read the task printed on paper, which represented the information need, (2) indicate the quantification type, (3) input the keywords to the search machine, (4) inspect the search result, write down the number of relevant documents and copy the pie chart, if any, on paper and (5) fill in the questionnaire.

Results

The four dependent variables were analyzed separately with repeated measures for analysis of variance (GLM). The cohesion between the dependent variables was analysed using linear regression and independent t-tests. We also collected the users' remarks, but we will concentrate here on the summary statistics.

Table 2: Means of the dependent variables in the two conditions *fulfilled* and *unfulfilled* and their levels of significance and F-values.

	Fulfilled (mean)	**Unfulfilled** (mean)	**Sig.**	**F**
Satisfaction documents	9.2	4.2	.000	463.62
Satisfaction search engine	9.0	4.2	.000	387.76
Continue to search	.29	.84	.000	153.40
Chart is useful	6.1	6.9	.072	3.48
Chart might be useful	6.0	7.4	.004	9.86

The influence of fulfillment on the dependent variables is clearly demonstrated in table 2. According to the significance levels, **H1** is confirmed except for the usefulness of the chart.

To avoid a tedious enumeration, we will focus on the main results now. So, instead of giving all the tables for all interactions, we will give table 3 as an example of what the data look like, and then summarize the others (for the reader who wants to study the details, we would be happy to make all the data available).

First we will look at **H2** and **H7**, concerning the intuitions of participants about recall feedback. In the condition unfulfilled, low recall indeed leads to different usefulness (F= 3.81, p= .034), satisfaction with the documents (F= 4.233, p= .013) and search engine (F= 6.803, p= .011). In the condition fulfilled, high recall leads only in type 2 tasks to different usefulness (F= 7.788, p= .007) and satisfaction (documents: F= 11.703, p< .001; search engine: F= 10.067, p= .002). The bahavior, however, is not influenced. This means that intuitions of participants do play a role in evaluation, but not in their subsequent behavior.

Table 3: Satisfaction with the documents, when the information need is fulfilled. 'Q Type' is the quantification type, 'Feedback' the recall feedback. The numbers indicate mean scores on the 11-point scale for user satisfaction.

Feedback / **Q Type**	absent	low	middle	high	overall
1	9.7	10.1	9.7	9.5	9.8
2	9.5	9.0	9.2	9.9	9.4
3	8.9	8.2	8.6	8.3	8.5

Now let's consider **H3-6** (fulfilled condition). The magnitude of the recall did not influence any of the variables. The satisfaction with both the documents and search engine was high but for type 3 lower than for type 1 and 2 (documents: 1-2: p= .276; 1-3: p< .001 and 2-3: p= .002; search engine: 1-2: p= .133; 1-3: p< .001 and 2-3: p= .005). **H3** and **H4** are therefore partly confirmed. As mentioned before, some participants do not agree with us that five relevant documents is enough to fulfill an information need of type 3. As a result, many participants want to continue to search in type 3 tasks of the condition fulfilled (34.4%). Also, in type 2 tasks of this condition many participants want to continue to search (44.5%). This can be explained by the restriction to *ten* documents in our experiment; it is impossible that these always include *all* existing relevant documents. In type 1 tasks, however, 93.0% want to stop searching; most participants obviously fulfilled their information need. Low recall did not cause a larger proportion of participants wanting to stop searching. **H5** is rejected because of these results.

The usefulness of the chart was not as low as expected, but did increase in order of type 1, 3 and 2, confirming **H6**.

Now I will discuss **H8-11** (unfulfilled condition). The satisfaction with both the documents and the search engine was low. Quantification type didn't influence them (p= .397 and p= .512). The satisfaction with the documents was influenced by recall (F= 4.233, p= .013) and was higher in absence of a chart, then in presence. But the satisfaction with the search engine was only in type 2 tasks influenced by the recall feedback low recall causes then a lower satisfaction than average recall (F= 6.803, p= .011), high recall (F= 11.449, p= .001) or no chart (F= 5.666, p= .020). **H8** is just partly confirmed and **H9** is rejected. Participants did want to continue to search (82.3%), confirming **H10**. The usefulness of the chart was not as high as expected, there was an effect

of type (F= 11.07, $p< .001$); it was highest for type 2, confirming **H11**.

There was a strong cohesion among the variables. Satisfaction with documents and search engine correlate strongly ($\beta = .92, p<.001$), Satisfaction with the documents correlates negatively with the usefulness of the chart ($\beta = -.102, p = .005$), and similarly for the estimated value of the chart, if it was absent ($\beta = .29, p<.001$). Similar values hold for the satisfaction with the search engine. The tendency to continue to search was strongly related to the other three dependent variables.

Discussion and conclusion

Quantification type did influence the results, contrary to our expectations. There are several plausible explanations for this influence. First, in several tasks, some participants did not agree with us about the number of relevant documents among the documents returned. We know that some uncertainty about the relevance of documents existed as some of the participants marked a document as relevant, that we found irrelevant and vice versa. We noticed this uncertainty because the numbers of the documents written on paper did not have complete overlap with the documents we found relevant. Second, there were varying interpretations of the phrase 'several documents' that we used to indicate type 3. Our pilot study indicated that 'several' could stand for 'about five' relevant documents but the participants of our experiment needed more relevant documents to be satisfied. Tasks that were meant to be fulfilled, may, in the eyes of some participants, not have been completely fulfilled and the other way around. Especially for quantification type 3 the satisfaction in the condition fulfilled was lower than expected. The large standard errors, especially in type 1 tasks, also reflect the differences among participants concerning relevancy and fulfillment. Hence, both the conditions fulfilled and unfulfilled are not as unequivocal as expected. Though judgments on relevancy are by definition subjective, more pilot studies could have increased the certainty in interpreting the results for both experimenter and participant.

Nevertheless, the experiment showed us that different types of information needs can be discerned. If search engines can get information about the type of the user's information need, they could adapt the exactness of its search and influence recall and precision (see figure 1)

Second, the results indicate that first, if participants are highly satisfied with the documents, they want to stop searching and they are not interested in the chart, and second, if participants are unsatisfied, they want to continue searching and understand that the chart provides worthwhile information.

To summarize. It seems that fulfillment is a better predictor of user satisfaction and behavior than precision and recall as used to evaluate IR systems. Search results with low precision can indeed result in high satisfaction, e.g. in case of quantification type 1.

Let is briefly comment on the question whether it is worth the effort to see if recall, which is a valid metric to compare the quality of IR systems, is also a valid metric for IR quality as perceived by the end-user. The participants in our experiment found the chart quite useful. This puzzles us, as it was meant to represent recall, and recall had very little overal effect. It might be that participants needed more time to familiarize themselves with the concept of recall feedback. The result is paradoxical enough to warrant further research. We stay with our prediction that recall information indeed will help the searcher in certain cases. But as long as a compromise must be found between recall and precision, precision should get a higher priority; the results suggest that even if the recall is low, the satisfaction can be high.

It is worth the effort to investigate ways to compute recall more precisely than can currently be done (e.g. pseudo recall or relative recall). The present authors are investigating a 'capture-mark-recapture' technique borrowed from biology, to observe in what proportion documents reappear in a search. In addition, we found a few cases where intuition conflicts with experimental findings. These may also be a source for further investigation.

Summarizing the main conclusions: First, among the variables we investigated, the one with the dominant influence on user satisfaction was whether the information need was fulfilled, and second, recall had virtually no influence on satisfaction or search behavior.

References

Buckley, C., & Voorhees, E. M. (2000). Evaluating evalutation measure stability. In *Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 33–40). New York: ACM.

Cleverdon, C., Mills, J., & Keen, M. (1966). *Factors Determining the Performance of Indexing Systems* (Tech. Rep.). ASLIB Cranfield Research Project.

Cooper, W. (1968). Expected search length: A single measure of retrieval effectiveness based on weak ordering action of retrieval systems. *Journal of the American Society for Information Science*, *19*, 30–41.

Hersch, W., Turpin, A., Price, S., Chan, B., Kraemer, D., Sacherek, L., & Olson, D. (2000). Do batch and user evaluations give the same results? In *Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 17–24). New York: ACM.

Hoenkamp, E., & de Groot, R. (2000). Finding relevant passages using noun-noun components. In M. Hearst, F. Gey, & R. Tong (Eds.), *Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 385–387). New York: ACM.

Hoenkamp, E., Stegeman, O., & Schomaker, L. (1999). Supporting content retrieval from WWW via 'basic level catogories'. In N. J. Belkin, P. Ingwersen, & M.-K. Leong (Eds.), *Proceedings of the 22rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 301–302). New York: ACM.

Swets, J. A. (1963). Information retrieval systems. *Science*, *141*, 245–250.

Modelling Language Acquisition: Grammar from the Lexicon?

Steve R. Howell (showell@hypatia.psychology.mcmaster.ca)
Department of Psychology, McMaster University, 1280 Main Street West, Hamilton, ON Canada

Suzanna Becker (becker@mcmaster.ca)
Department of Psychology, McMaster University, 1280 Main Street West, Hamilton, ON Canada

Abstract

A neural network model of language acquisition is introduced, based on and motivated by current research in psychology and linguistics. It includes both semantic-feature representations of words and localist linguistic representations of words. The network learns to associate the semantic features of words to their linguistic labels, as well as to predict the next word in the corpus. This is interpreted to model both the acquisition of a lexicon, and the beginnings of syntax or grammar (word order). The relationship of lexical learning to grammar learning is examined, and similarities to the human data found. The results may provide support for the 'Grammar from the Lexicon', or 'emergent grammar' position.

Introduction

How do children acquire language? More generally, how does any abstract language learner acquire language? When we attempt to model language processing via computer simulation, what should we be attempting to model, mature adult performance, or the developmental schedule of a child? What can such a model hope to tell us about the process of language acquisition in human infants?

These are some of the questions motivating our effort to model language processing. Much evidence exists as to the usefulness of the connectionist modelling enterprise for the understanding of human language in general. However, as we seek to model more fully the actual processing, and even production, of language, in a behavioural fashion, we consider it very important to take a developmental approach to human language processing. That is, a complete model of language processing should first become a model of language acquisition. Evidence suggests that a model of language acquisition in children should provide the foundation necessary to scale up to a model of more mature language processing, as we shall see.

Developmental Language Acquisition

In considering a developmental model of language, one important aspect is the limits of the enterprise. That is, where does language acquisition start, and where does it end? Language is a very complex cognitive activity, and our connectionist modelling techniques still maturing. We do not want to include any more than absolutely necessary in a model of language if we are to be successful. Thus, it is important to be explicit about our assumptions, in terms of pre-linguistic mental representations, or of what we can exclude from our model or include only as inputs.

We assume here that modelling any of the low-level acoustic properties of language is unnecessary for our purposes. While issues such as phonemic segmentation are important for language, those auditory tasks are arguably well-learned by the time of vocabulary acquisition. Further, modelling to the level of acoustics is too computationally demanding to include in a model of language acquisition at present.

If we consider the start of vocabulary acquisition to be at the age of the child's first word, typically 8-12 months, then we can ask the following question. What cognitive capacities does the child have prior to that point? What does language have to build upon? Some suggest that there is a considerable amount.

Lakoff and colleagues (Lakoff, 1986; Lakoff & Johnson, 1999) suggest that the child has reached an adequate level of concept formation prior to the development of language. Few would argue, we believe, that pre-linguistic children must have some kind of internal representation of the world, some understanding that a cat is fuzzy and can be patted, even if they don't know the words cat, or pat, or fuzzy. Lakoff argues that children's sensorimotor experience is continually building up these pre-linguistic concepts, concepts that are very specific and concrete, and that these concepts enable the child to function in their limited world.

With all of this cognitive machinery already well established, the language learning problem has happily become much simpler. If a child already has a concept for things like 'cat', then when it begins to learn the word for cat, it is really only attaching a linguistic label to a category of sensorimotor experience that it has previously built up. The learning of words is thus reduced to the learning of labels for things. The attributes of those things and the relationships between them are all predetermined (at least at this stage) by the child's environmental experience. Of course, nouns fit into this viewpoint with greater ease than do verbs; it is harder to point to a verb than a noun.

This is the traditional view in developmental psycholinguistics according to Gillette et al. (Gillette, Gleitman, Gleitman, & Lederer, 1999). As they point out however, this view has limits. Specifically, they show evidence that only some words can be derived solely via extralinguistic context.

It is well known that there is an overwhelming preponderance of nouns in children's early speech, not only in English but in most languages, while adults, of course, have a much more equal balance. Several explanations have been offered for this distinction. The discontinuity hypothesis holds that the cognitive capacities of children are fundamentally different from adults. Thus, at some point after the start of development of language children's cognitive capacity for language changes. Gentner describes the noun learning advantage as due to the conceptual complexity of the ways in which the two classes, noun and verb, describe the world (Cited in Gillette et al, 1999). That is, nouns describe object concepts, while verbs describe relations between objects. The latter would obviously be the more complicated task, since it depends on the success of the former. As Gillette et al point out, by this interpretation learning words is not just a matter of associating labels to concepts. Significant conceptual learning must occur as well. If true, this interpretation would argue against the conceptualization of language-age children as relatively conceptually stable, and would also invalidate one of the assumptions of our modelling approach.

Fortunately, Gillette et al. offer a different interpretation, the continuity hypothesis, which assumes that children are conceptually equipped to understand at least those concepts that underlie the words that adults typically use with them, both nouns and verbs. However, they argue that it is still possible to account for children's initial restriction to noun learning, using instead the different informational requirements of words that are necessary to uniquely identify them from extralinguistic context. They refer to their hypothesis as an information-based account, and describe several experiments that support this account.

Most importantly Gillette et al. provide strong evidence that learnability is not primarily based on lexical class. That is, it is not whether a word is a noun or a verb that determines if it can be learned solely from observation. Rather, they demonstrate that the real distinction is based upon the word's imageability or concreteness.

It is obvious that the very first words must be learned solely by the child attempting to discover contingencies between sound categories and aspects of the world, over many different exemplars. Gillette et al. demonstrate that the very first words used by mothers to their children are the most straightforwardly observable ones, and that as a group, the nouns are in fact more observable than the verbs. Furthermore, the imageability of a word is more important than the lexical class. The most observable verbs are learned before the less observable initial nouns, accounting for the few rare early verbs in children's vocabularies.

So, imageability or concreteness is the most important aspect of the early words, nouns and verbs alike, and it determines the order in which they tend to be learned by children. This result argues against the discontinuity hypothesis, and supports Lakoff's early concepts and the borders that we have drawn for our language modelling enterprise. However, what of the less imageable words? How are they learned?

Gillette et al. also find evidence for the successive importance of noun co-occurrence information and then argument structure. That is, for later learning of the less imageable words (mostly verbs), observing which previously known nouns co-occur in a sentence with the yet unknown word label helps greatly to uniquely identify the concept. Thus rather than imageability determining exactly which object we are talking about over multiple experiences, for many verbs the nouns involved act to identify it. Thus if the noun 'ball' is paired with a yet unknown word, the concept 'throwing' may be activated for many learners, allowing them to infer that the unknown word means 'to throw' (Gillette et al, 1999). Argument structure is yet a further step to verb inference. Gillette et al. show that the number and position of nouns in the speech stream reliably cues which verb concept the unknown word could be.

At this point in the child's language learning we have moved beyond initial lexical learning and are in the realm of syntax. The first words (mainly nouns) have been learned without reference to other words, their sheer imageability enabling them to be inferred from the adult to child speech stream and the extralinguistic evidence. The next step involves the use of these concrete nouns to help infer the less imageable verb meanings in the speech stream, and from there the child is no longer learning words solely from the extralinguistic context. The lexical structure of utterances now assists the child as well. For example, the first few verbs learned, when experienced in adult speech and involving a novel object, will cue the inference of the new noun label and, depending on the particular verb, even the type of noun involved. The circular, bootstrapping process of language learning is on its way (for further evidence concerning verbs and nouns respectively, see Goldberg, 1999; Smith, 1999). Before long new words will no longer require explicit extralinguistic context at all. The school-age child will begin reading and acquiring new words solely by lexical constraints, allowing them to exhibit the incredible word acquisition rates that have been reported (e.g. Bates & Goodman, 1999).

Of course, once the child's lexicon has reached a certain level of complexity, perhaps 300 words (Bates and Goodman, 1999) the multi-word stage begins, and grammar acquisition begins to be a consideration as well as just lexical acquisition.

Grammar From the Lexicon

Bates and Goodman (1999) examine the highly linked development of grammar and the lexicon. They provide evidence for the emergence of grammar directly from the lexicon itself. Specifically, they show the lack of evidence for any dissociation of lexical and grammatical processes (drawn from studies of early and late talkers, focal brain lesions, and developmental deficits), along with the very tight developmental ties between the two. For example, lexical status at twenty months (during children's vocabulary burst) is the single best predictor of grammatical status at 28 months (during children's grammar burst), with a correlation coefficient of between .70 and .84. This is in fact as good a statistical relationship as that between separate measures of grammar! This is good evidence that grammar does emerge, at least partially, from the very growth of the lexicon itself.

This finding, as well as those of Gillette et al, is important to the development of our model of language acquisition, as if grammar development is emergent from lexical development, then we want to be sure that we do not model them as two separate modules or components. Rather, a central tenet of our model is to use a single processor architecture to learn both lexicon and grammar. Furthermore, lexical development should precede grammatical, and grammatical development should not take off until sufficient lexical development has occurred. Our model should exhibit the same sort of acquisition (and production, eventually) behaviour as a child.

A Dynamical Systems Approach

Elman (1995) suggests viewing the process of initial lexical and grammatical development as a dynamical system, or attractor model, which can be learned through a process of predicting the input. Roughly speaking, this viewpoint is as follows. A language learner's semantic representations are very limited at first, much like a flat three-dimensional landscape. Then as the learner develops stable categories and concepts, the landscape gradually develops depressions or basins, each basin corresponding to a word or concept, and each experience of that concept deepening the basin, until eventually the landscape is full of deep and wide basins of attraction. These are "attractors" since, while any partial or confused activation of a semantic representation will tend to indicate a place on the landscape not in one of these basins, the slope of the 'terrain' is such that the representation will tend to be drawn down into one basin or another, and the larger basins will be more likely to capture the activation. They "attract" the activation.

Furthermore, this attractor representation is hierarchical. General or superordinate concepts might have very large basins, containing within them smaller basins corresponding to more specific but semantically related terms.

Obviously this landscape representation only applies to the lexicon. How does grammar enter into the picture? Well, if the lexicon is viewed as basins in this representation landscape, or state-space, then grammar is contained in the transitions that occur between these states. That is, a true dynamical system consists not only of these representations in state space, but also relationships that influence movement from one representation to another. Further details can be found in Elman (1995), but for our present purposes it is sufficient to realize that this dynamical systems approach provides a possible mechanism for the implementation of the word-inference processes described earlier (Gillette et al. (1999). Certainly a recurrent net like the one we will describe in our model is capable of exhibiting the behaviour of a dynamical system, with the hidden unit representations corresponding to the state-space vectors and the operation of the network providing the transitions between them based on the values stored in its weights. It can also be argued that the cortex operates in this fashion (Elman, 1995; Sulis, 2001, personal communication), and thus that the same explanation can be offered for human language processing.

The 'Complete' Early Language Acquirer

Let us assume, then, that the child (or model) starts with pre-existing pre-linguistic concepts of the world, upon which linguistic labels will be learned by direct instruction as well as simple exposure. This pre-existing conceptual structure implies either a pre-existing mental representation (semantic landscape) or one that is quickly built up as words are matched to concepts.

In our model, we assume that the child begins syntax or grammar learning at the same time as it begins learning vocabulary. However, since there is little evidence that grammar is directly instructed (Bates & Goodman, 1999), unlike noun acquisition (Smith, 1999), and since grammar is inherently more complex, grammar learning does not really succeed until after the most primal of the lexical attractors have been firmly set and the lexical and syntactical bootstrapping has begun. In essence, grammar exposure begins at the same time as lexical learning, but grammar learning doesn't effectively take place until the lexical representations are solidified.

Thus we would expect to see exactly that behaviour that is seen in real children; lexical development proceeds at an ever accelerating pace, then when the lexical foundation is firm enough (the 'noise' or uncertainty in the language environment is reduced enough) the mental machinery can focus on syntactic relationships, and grammatical learning should accelerate. Our model should exhibit exactly this behaviour if it is capturing the essence of human language acquisition.

Method

Our experiment consists of training our model of language acquisition many times from different initial conditions, and analyzing the performance results for their fit to the human data and improvements over the control models.

The Model

The model of language acquisition discussed herein (see Figure 1) takes as input uniquely identified words (localist input representations), and learns how those words can be used in sentences. This is not a novel undertaking (see Elman, 1990, 1993; Howell & Becker, 2000). However, what is new to this model is the addition of a second set of inputs, semantic-feature inputs. By 'semantic', however we actually mean pre-linguistic semantics or meaning (e.g. sensorimotor features). Thus, instead of abstractly manipulating locally-distributed word representations, a process that has been characterized by McClelland as "learning a language by listening to the radio" (Elman, 1990), our model attempts to ground the word representations in reality by associating them with a set of these semantic features for each word.

Furthermore, the network is not performing only the prediction task that is argued (Elman, 1990) to lead to an internalization of basic aspects of grammar, specifically word-order relationships. Instead, it is also learning, simultaneously, to memorize its linguistic inputs, memorize its semantic inputs, and associate the two together, such that either one alone will elicit the other.

Why construct a neural network model in this way? First, using a simple recurrent architecture and prediction task retains the successful grammar learning capabilities that have been shown so well by Elman and colleagues. Second, adding a semantic layer will eventually allow for the use of phonemic input representations and the binding of those phonemes into words (through semantic constancy across each individual word) although the network discussed in this paper does not deal with phonemic inputs, only whole-word inputs. Third, the inclusion of the semantic input layer and a semantic output layer means that semantic features can be read off for any given linguistic input, indicating whether the network has learned the "meaning" of the word.

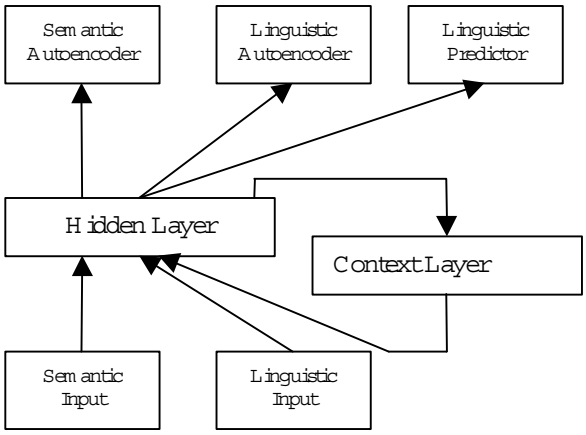


Figure 1: Modified SRN architecture, including standard SRN hidden layer and context layer, standard linguistic prediction layer, and novel semantic autoencoder and linguistic autoencoder.

Finally, the inclusion of both linguistic autoencoding (word learning) and linguistic prediction (grammar learning) allows us to explore the dynamics of the model, and determine if the learning behaviour of the model maps to the human developmental data. That is, does the word learning have to reach a critical mass before the grammar learning proceeds? Does a jump in lexical competence lead to a linked jump in grammatical competence? If so, then perhaps the model can provide evidence for the view that grammar emerges from the lexicon (Bates and Goodman, 1999).

Model Details

There are two input layers and three output layers. The semantic output layer is paired with the semantic input layer. Both are 68 nodes in size, since the semantic feature dimensions taken from Hinton & Shallice (1991) have 68 dimensions.

The linguistic input and the linguistic outputs are of size 29, since the vocabulary has 29 words. Both linguistic outputs are tied to the same set of linguistic inputs, but where the linguistic autoencoder's training signal is the present input, the linguistic predictor's training signal is the input at the next time step.

Both the hidden and the context layer are of size 75, and the hidden-to-context transfer function is a one-to-one copy with no hysteresis (see Howell & Becker, 2000). The hidden-to-context connection is not

trainable, but the context-to-hidden feedback connection is trained exactly as is either of the input-to-hidden connections.

Training Environment

The network is trained on a corpus of text derived from a small (390 word) subset of Elman's original corpus of two and three word sentences with a 29 word vocabulary (Elman, 1990).

Input to the semantic input layer was derived from the above corpus by converting each word in the corpus to the word's semantic featural representation, using a set of features derived from Hinton and Shallice (1991). This feature set includes only the sensory features and excludes the semantic-association ones found in the original. This resulted in a binary distributed representation for the semantic layer. It is important to note that on language tasks a binary distributed representation would often be expected to learn faster than a localist representation, as it provides more information to the network.

The network's weights were randomly initialized, and training proceeded as usual for Simple Recurrent Networks, using the backpropagation algorithm (Rumelhart, Hinton, and Williams, 1986).

Training proceeded until reasonable levels of accuracy were achieved. Trial runs of up to 1500 epochs indicated that the net asymptoted near 500 epochs, so training did not in any case proceed beyond 500 epochs.

Error measures and accuracy measures were logged at each input presentation, but averaged over the 390 patterns to one value per epoch of training.

Results & Discussion

The first finding from the various runs of the network is that the net does in fact learn. There had been some concern that the demands of three different tasks sharing a single hidden layer might cause significant or even catastrophic interference in the learning tasks. On the contrary, with a hidden layer size only slightly larger than the largest input layer (75 compared to 68 for the semantic input layer) the net learned all three tasks.

Furthermore, the tasks were learned in the expected order. That is, judging from the error curves the binary distributed semantic representations were learned most quickly (since they provide more information for the network to learn on, i.e. more bits turned on) followed by the localist linguistic autoencoding and then the localist linguistic prediction. Prediction, of course, is a more difficult task than autoencoding or 'memorization', just as verb learning is a more difficult task than noun learning.

For the present purposes, our analysis is limited to the lexical-grammatical relationship (and further semantic results are not reported). Specifically, over 24 simulation runs the mean peak lexical accuracy was 96.6 percent correct, while the mean peak grammatical accuracy was 37.33 percent correct (See Figure 2).

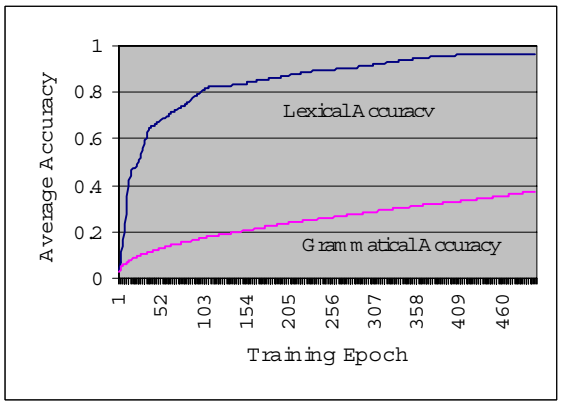


Figure 2: Average Accuracy Curves Over 24 Runs

Comparisons with 'control' or partial networks lacking the semantic or lexical autoencoder task also indicate that each task is learned faster and more accurately in the experimental network than in the control networks. Only the grammatical results are reported here, however.

For control network 1, which included only the linguistic prediction task (i.e. an original Elman net) the peak prediction accuracy was lowest, with a mean of 18.5 percent correct, and significantly different from the experimental network via t-test ($n = 10, p < 0.0001$).

For control network 2, which excluded only the semantic layers, the peak prediction accuracy, achieved at epoch 500, was significantly lower than the experimental network ($n=10, m = 28.4, p < 0.0001$).

For control network 3, which excluded only the linguistic autoencoder, the peak prediction accuracy was still lower than the experimental network ($m = 37.1$) but the difference did not reach significance ($n=10, p = 0.137$).

Thus, training all three tasks through a single hidden layer apparently creates synergies that allow each to proceed faster than it would alone.

Most interesting, however, was the relationship between the lexical and grammatical accuracy curves for the experimental network. We expect that if our model is catching important elements of the human language learning experience, then it should exhibit lexicon-then-grammar behavior. Certainly, as discussed above, the speed of learning (rate of error decline) exhibits this relationship, but that is only to be expected by the difficulty of the tasks. A better question is

whether the network exhibits the lexical-to-grammatical performance correlations that Bates and Goodman (1999) discuss. That is, does the lexical performance at time T correlate well with the grammatical performance at some later point?

By analogy to the methods cited in Bates and Goodman (1999), a point on the lexical accuracy curves that could be considered the 'lexical burst' was identified (approx. Epoch 108). Then, since there was no explicit 'grammar burst' within our time window a set of correlations was calculated to the grammatical performance at various time lags from the lexical burst (see Figure 3). The results indicate that the highest correlation, approximately .80, is from the lexical burst to grammatical performance 75 epochs later.

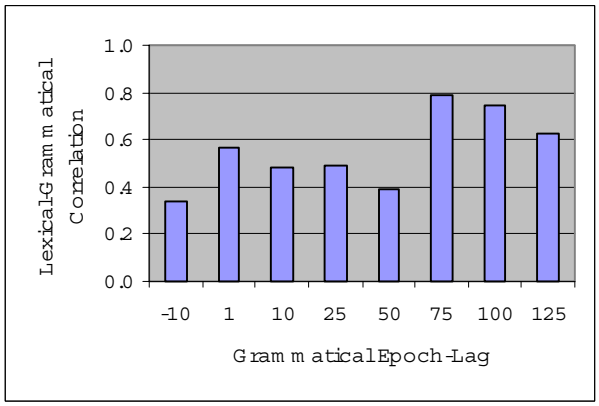


Figure 3: Lexical-Grammatical Correlations (n = 24)

This is similar to Bates & Goodman's cited correlation between lexical status at 20 months and grammatical status at 28 months in children. At first, the similarity may seem limited, since our model uses only 29 words, not the 300-plus that is suggested to be the critical mass required for grammar learning. Also, our sentences use only the 29 words from the model's vocabulary, and no unfamiliar words, and word learning is being represented by average accuracy curves. Further, grammatical status is being measured by accuracy of prediction rather than Mean Length of Utterance (MLU).

However, we believe these results are promising, and that further study is warranted. We have already begun to run simulations that use larger vocabularies, and that provide analogues of MLU measurements for grammatical status, in order to elucidate further the model's relationship to human performance.

Acknowledgments

Thanks to George Lakoff, whose writings and personal conversations inspired some of this work, and to Damian Jankowicz, whose comments were most helpful throughout. This work has been supported by a Post-graduate Fellowship from the National Sciences and Engineering Research Council of Canada (NSERC) to SRH, and a research grant from NSERC to SB.

References

Bates, E. and Goodman, J. C. (1999). On the emergence of grammar from the lexicon. In MacWhinney, B. (Ed.) (1999). The Emergence of Language. New Jersey: Lawrence Erlbaum Associates.

Elman, J. L. (1990). Finding structure in time. Cognitive Science, 14, 179-211.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. Cognition, 48. 71-99.

Elman, J. L. (1995). Language as a dynamical system. In R. F. Port & T. van Gelder (Eds.), Mind as Motion: Explorations in the Dynamics of Cognition. Cambridge, MA: MIT Press.

Gillette, J., Gleitman, H., Gleitman, L., Lederer, A. (1999). Human simulations of vocabulary learning. Cognition, 73, 135-176.

Goldberg, A. E. (1999). The emergence of the semantics of argument structure constructions. In MacWhinney, B. (Ed.) (1999). The Emergence of Language. New Jersey: Lawrence Erlbaum Associates.

Hinton, G. E. & Shallice, T. (1991). Lesioning a connectionist network: Investigations of acquired dyslexia, Psychological Review, 98, 74-75.

Howell, S. R. & Becker, S. (2000). Modelling language acquisition at multiple temporal scales. Proceedings of the Cognitive Science Society, 2000, 1031.

Lakoff, G. and Johnson, M. (1999). Philosophy in the flesh: The embodied mind and its challenge to western thought. New York: Basic Books.

Lakoff, G. (1987). Women, fire and dangerous things: What categories reveal about the mind. Chicago and London: University of Chicago Press.

Rumelhart, D E., Hinton, G. E. & Williams, R. J. (1986) Learning internal representations by error propagation. In J. L. McClelland, D. E. Rumelhart and the PDP Research Group, Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations (pp. 318-362). Cambridge, MA: The MIT press.

Smith, L. B. (1999). Children's noun learning: How general learning processes make specialized learning mechanisms. In MacWhinney, B. (Ed.) (1999). The Emergence of Language. New Jersey: Lawrence Erlbaum Associates.

The Strategic Use of Memory for Frequency and Recency in Search Control

Andrew Howes (HowesA@cardiff.ac.uk)
School of Psychology, Cardiff University, Cardiff, CF10 3YG, Wales, UK

Stephen J. Payne (PayneS@cardiff.ac.uk)
School of Psychology, Cardiff University, Cardiff, CF10 3YG, Wales, UK

Abstract

A requirement of an information processing account of human problem solving is that it includes a mechanism by which people remember which goals and operators have been evaluated and which still need to be evaluated. One might expect that these are issues of such fundamental importance that they must have been solved or at least addressed by the two architectural accounts of cognition (Soar and ACT-R), but in fact it is an issue that is glossed in both. We identify two problems: (1) Soar and ACT-R guarantee information about goals, and (2) ACT-R combines measures of frequency and recency into a single representation of activation. In this paper we report a model of how people search simple binary trees. The model demonstrates the cognitive plausibility of a search algorithm that is supported by a memory system that delivers independent estimates of frequency and recency.

Introduction

A requirement of an information processing account of human problem solving is that it includes a mechanism by which people remember which goals and operators have been evaluated and which still need to be evaluated. Whether the task is the Tower of Hanoii, a waterjugs problem, a world-wide web search problem or a spatial navigation task, a person engaged in search examines the consequences of applying an operator to a state by trying it out and perceiving to which state it, and subsequent operators, lead. At some point in the future, the person may, through backup, or because of loops, find themselves in a visited state. Recognition that the state has already been visited and/or that the operator has already been applied to this state, will in the long-term help prune the search space and thereby constrain the effort spent on attaining the goal. This constraint has been used in a number of models of human problem solving (Atwood & Polson, 1976; Jeffries, Polson., Razran, & Atwood, 1977; Anderson, 1993; Howes, 1994). Atwood & Polson's model of human performance on the waterjugs problem, built up a representation of the 'familiarity' of states that was factored into the operator selection process. The more familiar an operator then the less likely it was to be selected.

One might expect that these are issues of such fundamental importance that they must have been solved or at least addressed by the two substantial architectural accounts of cognition (ACT-R, Anderson, 1998; Soar, Newell, 1990), but in fact it is an issue that is glossed in both. In Soar, the architecture automatically ensures that operators that have already been applied to a particular state in pursuit of a particular goal (on the goal stack) on a particular trial will not be reselected. In ACT-R the goal stack has privileged status. Items posted on the stack are not subject to the constraints of memory, i.e. they do not have decaying activation and cannot therefore be forgotten (Altman and Trafton, 1999).

Another resource for supporting decisions about which operator to apply is memory for previous attempts at a goal (either successful or failed). If a goal has been achieved prior to the current attempt then memories that indicate that an operator is familiar may be taken as evidence that it is more likely to lead to the goal than an unfamiliar operator (Payne, Richardson, Howes, 2000). However, an issue for the problem solver is how to determine the source of the familiarity. If the source is the current trial then the operator should be rejected, if it is a previous trial then perhaps it should be selected.

Payne, Richardson, Howes (2000) investigated the role of familiarity (Jacoby, 1991) in controlling interactive search. They tested the hypothesis (Aasman & Akyurek, 1992; Howes, 1994) that people help control search merely by recognising the actions that have been tried before and found that the familiarity of items could affect decisions about which item to select. Moreover familiarity was used strategically. When participants had information indicating that familiarity would be more likely to indicate that an operator would lead to the goal, they were more likely to use familiarity to guide selection.

Again, one might expect that this issue would have been addressed in architectural theories of cognition. However, while Soar's chunking mechanism is flexible, the issue of whether it can provide a mechanism for representing the episodic familiarity of an operator has only recently started to be explored (Altmann and John, 1999). The situation for ACT-R is more complex.

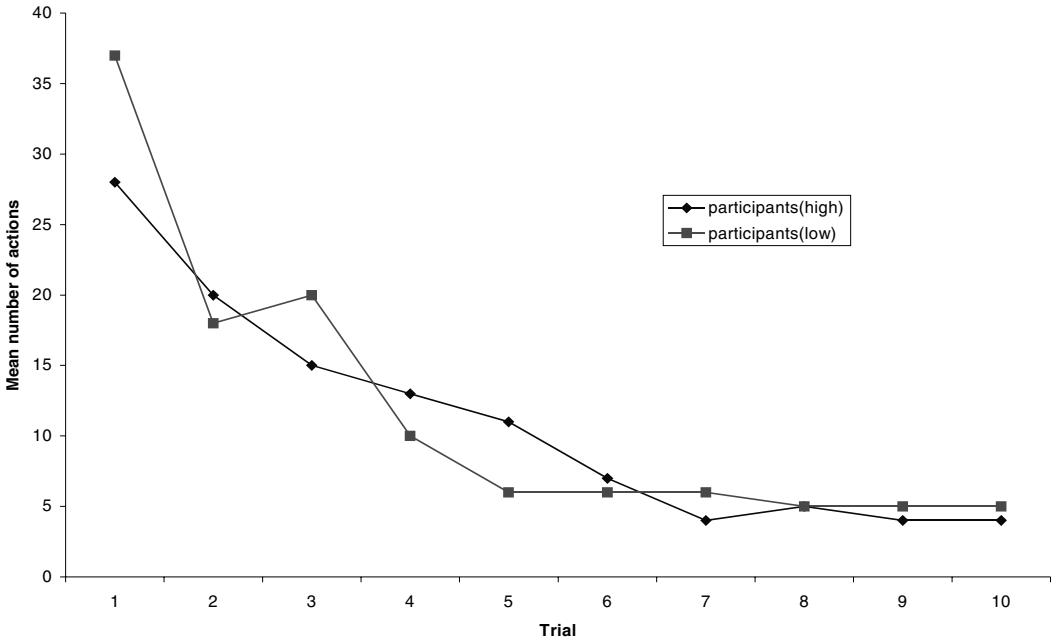


Figure 1: Mean number of actions taken by high and low systematicity participants

In ACT-R, each chunk stored in declarative memory has an activation that is used to determine probability of retrieval. This activation is made up of a base-level activation and an associative activation. Anderson and Lebiere (1998; page 70) state: "... the activation of a chunk is a sum of the base-level activation, reflecting its general usefulness in the past, and an associative activation, reflecting its relevance to the current context," and, "The base-level activation of a chunk represents how recently and frequently it is accessed."

Importantly however, the frequency and recency components of base-level activation are not independently inspectable by the production rules and it is not therefore possible to write ACT-R production rules that make strategic use of frequency and recency information stored as components of chunk activations. It seems unlikely therefore that it is possible to write productions that, for example, prefer the most frequent operators at the expense of the most recent.

A commonly used solution to this in ACT-R models has been to use flags on declarative memory structures. A flag is added to operators that have been applied on this trial and then all flags are wiped at the end of the trial (Anderson's model of navigation 1993; Lebiere, personal communication) leaving no episodic evidence that they had ever been there. While, this is an extra-architectural mechanism that, unsurprisingly, is not claimed as part of the theory, its use undermines the claim that models constructed in ACT-R are subject to a principled set of memory constraints.

In this paper we report a computational level model of how people search simple binary trees. The model makes strategic use of frequency and recency information and demonstrates the cognitive plausibility of a search algorithm that is supported by a memory system that delivers independent estimates of frequency and recency.

Task

In a series of studies to be reported elsewhere we observed participants searching simple binary word-mazes. Each maze was a binary tree structure with a depth of 5 nodes. At each choice point participants were presented with three buttons on a computer display. Two buttons at the top and bottom of the right of the screen were labelled with different words (perhaps 'gun' and 'pistol') and the other button, on the left of the screen, was labelled 'back'. Selection of one of the two buttons on the right changed the current state to a state nearer to the leaves of the tree and selection of 'back' moved the state to a node nearer the root of the tree. Participants were asked to search for a leaf node with a given label (a random word).

Observations:

- All participants were able to complete these search tasks.
- Three strategies were used:
  - **Random search** with a forward bias. Participants selected either the top or the bottom button on the right of node X, searched the subtree and then on returning to X selected the other button.

- **Systematic search**. Participants always selected the top button on the first visit to node X, searched the subtree, and then on return to X selected the bottom button.
- **Memory-based search**. On trials after the first participants generally attempted to remember the correct path.

- On trials after the first, participants flexibly interleaved search based on memory for previous trials with, when memory for previous trials failed, either systematic or random search.
- None of the participants perseverated, i.e. they did not repeatedly search the same incorrect subtree more than a handful of times.
- With practice (about 4 trials) all participants were able to follow the correct path with relatively few errors (Figure 1).
- Those participants who used a systematic strategy were significantly more efficient than those who did not. On the first trial the variation in the performance of the systematic participants was less than the variation in the performance of the random participants. (Unfortunately, the statistically significant difference in efficiency between the use of the two strategies is not reflected in Figure 1. This is because some participants using the random strategy can, luckily, find the goal with relatively few actions.)

Model

The first model that we built relied on a single activation-based measure that combined both frequency and recency information. This model could perform the first trial of a task by avoiding operators with high activation (those inferred to have been selected recently or frequently). However, on subsequent trials, a strategy of preferring operators with a higher activation (i.e. the ones used most recently on the previous trial or the ones used most frequently over trials) proved to be fragile. Activation may be high either because an operator was selected many times incorrectly or because it was selected more recently (i.e. closer to the achievement of the goal). Worse, if an error is made because an algorithm prefers highly active operators, then the algorithm may perseverate ad infinitum on incorrect selections.

The model that we focus on in this paper, is an extension of a proposal by Payne, Richardson, Howes (2000). It relies on the separate and strategic use of information about the frequency and recency of operator usage. The model is not based on assumptions about the structure of memory, rather it is based on assumptions about what information memory can deliver. The heuristics that define the search algorithm rely on the following functions for acquiring information from memory:

- **F = frequency(I)** - returns an estimate of the frequency F of item I.
- **X = most_recent(P)** - Instantiates pattern P to its most recent occurrence. (e.g. most_recent(op) would bind X to the most recently tried operator). Only one value can be returned for a particular P.
- **F = freq_before(I, E)** - returns the frequency F of I before the most recent occurence of event E. (e.g. to give the frequency of an item I before the selection of the current goal.)
- **F = freq_after(I, E)** - returns the frequency F of I after event E.

In order to simulate a lack of reliability in the information returned by these functions, frequency and recency information decayed from memory stochastically. Also, false positives were randomly generated in answer to queries about whether operators had been applied on this trial. In the Payne, Richardson, and Howes (2000) experiment, false positives occurred when participants were forced to make a decision about whether or not they had applied an operator before. In fact, participants may have only seen the operator and not applied it. The functions that determined the rate of decay and false positives are not important for our current purposes.

The purpose of introducing the errors was not to capture some quantitative aspect of the data but instead to ensure that the search algorithm was robust given the return of incorrect information from memory. Most importantly the algorithm should not perseverate implausibly even when degraded information is returned from memory.

The heuristics work by adding to a preference value for each operator proposed. There are three sets of heuristics: those that switch algorithm (or strategy); those that control systematic search; and those that control frequency-based search.

Given goal G, operator Op and a preference constant V, the rules for each algorithm are described below. The rules depend on memory encodings of the frequencies and recencies of associations, in general between G and Op, but for clarity. a short-hand has been used to describe the rule conditions, which does not refer to the association per se, but instead just to Op. Each rule proposes an addition (plus) or a subtraction (minus) to the current value of the preference for Op. The rules are described in a pseudo-code where variables are represented with capitals. The symbol '=' indicates a test of equality. If the test has a variable on either side and the variable is not already bound then the test will result in binding. The variables TOP and BOTTOM are respectively bound to the top and bottom forward menu selections.

Rules 1 to 5 describe the memory-based algorithm. This algorithm is used if the model has a memory

indicating that the goal has been achieved before. Rules 6 and 7 describe the random algorithm. Rules 8 to 10 describe the systematic algorithm. (A particular instantiation of the model uses either the random or the systematic rules but not both.) Finally, rule 11 switches to the memory algorithm and rule 12 restarts a search in the case of apparent exhaustion (this is described further below).

There is only space to describe some of these rules here. We will focus on those for the systematic algorithm. Rule 8 says, if the most recent algorithm is systematic and the operator (Op) being evaluated is a forward operator at the top of the screen, and the most recent of the previously applied operators (R) was not a 'back' operator THEN add V to the preference for Op. Rule 9 is similar to rule 8 but adds a preference for the forward operator at the *bottom* of the screen if the previous operator was a backup, Lastly, rule 10 prefers the back operator when the bottom operator has been tried on this trial and the most recent previous operator was also a back.

1. IF most_recent(algorithm) = A,
A = use_memory,
forward(Op) = true,
freq_after(Op,A) = 0,
freq_before(achieved(G), A) = FG,
freq_before(Op, A) = FO
THEN P becomes plus(1 / (1 + abs(FO – FG) * V))

2. IF most_recent(algorithm) = use_memory,
forward(Op) = true,
freq_before(fail(Op), now) = FN
THEN P becomes minus(FN * V).

3. IF most_recent(algorithm) = A,
A = use_memory,
forward(Op) = true,
freq_after(Op, A) = 0,
freq_before(Op, achieved(G)) = OF,
THEN P becomes plus(OF * V).

4. IF most_recent(algorithm)=A, A= use_memory,
forward(Op) = true,
freq_after(Op, A) = 0,
THEN P becomes plus(V).

5. IF most_recent(algorithm)=A, A= use_memory,
back(Op) = true,
freq_after(TOP, A) > 0,
freq_after(BOTTOM, A) > 0,
THEN P becomes plus(V).

6. IF most_recent(algorithm) = A, A = random,
forward(Op) = true,
freq_after(Op, A) = 0,
THEN P becomes plus(V).

7. IF most_recent(algorithm) = A, A = random,
back(Op) = true,
freq_after(TOP, A) > 0,
freq_after(BOTTOM, A) > 0,
THEN P becomes plus(V).

8. IF most_recent(algorithm) = systematic,
forward(Op) = true, top(Op) = true,
most_recent(op) = R, not(back(R) = true),
THEN P becomes plus(V).

9. IF most_recent(algorithm) = A, A = systematic,
forward(Op) = true, bottom(Op) = true,
most_recent(operator) = R,
back(R) = true,
freq_after(Op, A) = 0,
THEN P becomes plus(V).

10. IF most_recent(algorithm) = A, A = systematic,
back(Op) = true,
most_recent(operator) = R,
back(R) = true,
freq_after(BOTTOM, A) > 0,
THEN P becomes plus(V).

11. IF most_recent(algorithm) = A, A = none,
freq_before(achieved(G), A) > 0,
Op = algorithm(use_memory),
THEN P becomes plus(3*V).

12. IF current_node = root ,
most_recent(algorithm) = A ,
Op = algorithm(A),
freq_after(TOP, A) > 0,
freq_after(BOTTOM, A) > 0,
THEN P becomes plus(3*V).

The last algorithm switching rule (rule 12) plays a crucial role. Occasionally the problem solver will return to the root node without having found the goal. This will happen if the search was incomplete (i.e. some subtree remained unsearched) due to inadequate information from memory (a false positive). In this situation rule 12 restarts the search. In the model this is operationalised as the operator for the current algorithm is reapplied. The time at which the most recent algorithm operator was applied is used by the other rules to judge whether memories for operator applications were part of the current trial or previous trials.

Results

For particular rates of memory decay and false positives, the model was run 40 times on each of the 4 tasks performed by participants. The resulting mean performance for three decay rates is shown in Figure 2.

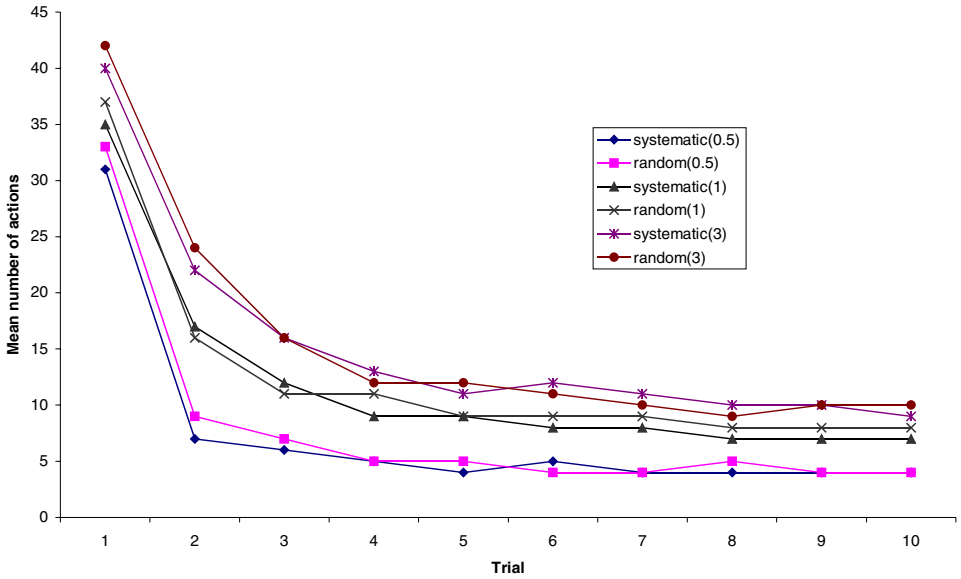


Figure 2: Mean number of actions taken by model given increasingly unreliable information from memory

The participants' mean performance is within the bounds of the best and worst model performance illustrated in Figure 2. We have not attempted to fit the model precisely, rather in accordance with Roberts and Pashler (2000) we explored the range of its behaviour.

Importantly, the model did not perseverate. Regardless of errors made during search, it always recovered and eventually found the goal. Also, as the decay rate increased the model was still able to learn the task. A large number of errors in the first trial did not on average incapacitate the learning over subsequent trials.

The gradual improvement in practice after the first trial was a result of a search algorithm (rules 1 to 5) that is guided by a combination for memory for previous trials and the current trial. If memory for previous trials proved inadequate then memory for the current trial, as distinguished by relative recency, ensured a reasonably efficient search.

Also, in accordance with the participants behaviour, the systematic algorithm produced more efficient and less varied searches on trial 1.

Discussion

The model reported here demonstrates that aspects of the way in which people search and learn paths through external problem spaces can be captured with heuristic rules that make strategic use of independent estimates of the frequency and recency of previously selected operators. Without access to this information it is impossible to write heuristics that distinguish an operator with high frequency from one that has high recency, and it is therefore a problem to determine whether key events occurred on the current trial or previous trials. The analysis of the model's behaviour under a range of memory decay and false positive conditions reveals that it produces behaviour broadly similar to human performance on a simple search task. Notably, unlike previous activation-based models built by the authors, the model does not perseverate when receiving degraded information from memory. In addition, the mean performance of the model over ten trials consists of a practice curve similar to that of the participants.

However, further investigation revealed that, after the first trial, the model produced a much greater variation in behaviour than the participants in the experiment. This issue is a matter for further investigation, and may well imply the need for some superordinate learning mechanism (perhaps rehearsal or impasse-driven learning).

A superordinate learning mechanism might involve the deliberate encodings of what the correct option is. This is an approach that was explored in Howes (1994), and while it deserves further attention, there are two problems. The first is that there is a dislocation in time between when the items are experienced and when a participant achieves the goal. In previous models the feedback of information about correctness produced recency effects in which lower levels of the tree were learnt first (Howes, 1994). These effects were not observed in our experiments. The second is that deliberate learning only pushes the problem back one level. If people deliberately learn what is correct then when situations change or mistakes are made, they also have to deliberately learn that a different option is

correct. Subsequent competition between different representations of correctness would then have to be resolved, perhaps using exactly the kind of mechanism that we have proposed. Progress will require modelling the range of individual trial data rather than just mean data.

Another possibility that we are investigating is that the long-term learning is based on recency and not frequency. Once the goal has been achieved, then the operator that led to the goal will be the most recently selected operator at any choice. Memory for recency could therefore be used to guide learning. However, under normal assumptions about decay, a recency-based model predicts that choice points at different distances from the goal would be learnt at different rates. Our data (not described above) does not support this prediction.

In principle, it may be possible to construct algorithms in ACT-R designed to ensure that during search sufficient episodic information is stored in declarative chunks to enable the kinds of computations that are posited in the model report here (e.g. Altmann and Trafton, 1999). However, regardless of the success of this approach, there will remain an issue about how people obtain information about frequency, and recency. While the concept of activation is well established in psychology, an architecture in which chunks are stored with independent measures of frequency and recency may lead to more parsimonious accounts of problem solving behaviour.

There are a number of models of the cognitive activity that give rise to practice effects, amongst them Logan's (1988) instance model and Rosenbloom and Newell's (1981) chunking model. More recent work has emphasised the strategy specific nature of the practice curves (Delaney, Reder, Staszewski & Ritter, 1998). The model reported here is similar to Logan's in that the practice curve emerges as a result of encodings made from experience with the external environment: however, maze-like tasks are more complex than simple letter arithmetic tasks and it is for this reason that our model requires the combination of frequency and recency dependent control mechanisms that we have described.

In the introduction we claimed that ACT-R's representation of undifferentiated activations was not sufficient to directly support algorithms that capture the behaviour of people engaged in typical search tasks. In contrast, the model that we have reported illustrates the cognitive plausibility of a mechanism that makes strategic use of separate sources of operator recency and frequency during search.

Acknowledgement

Juliet Richardson contributed a great deal to the development of the ideas presented in this paper.

References

Aasman, J. & Akyurek, A. (1992). Flattening goal hierarchies. In J.A.Michon & A.Akyurek (eds.) *Soar: A Cognitive Architecture in Perspective* , 199 -217. Kluwer.

Altmann, E.M. & John, B.E. (1999). Episodic indexing: A model of memory for attention events. *Cognitive Science, 23*(2),117-156.

Altmann, E. M. & Trafton, J. G. (1999). Memory for goals: An architectural perspective. *Proceedings of the twenty first annual conference of th e Cognitive Science Society* (pp. 19-24). Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. & Lebiere, C. (1998). *The Atomic Components of Thought*. NJ: Erlbaum.

Atwood, M. E. & Polson, P. G. (1976). A process model for water jug problems. *Cognitive Psychology, 8*, 191-216.

Delaney, P. F., Reder, L. M., Staszewski, J. J., & Ritter, F. E. (1998). The strategy-specific nature of improvement: The power law applies by strategy within task. *Psychological Science, 9*, 1-7.

Howes, A. (1994). A model of the acquisition of menu knowledge by exploration. In B. Adelson, S. Dumais, J. Olson (Eds.) *Proceedings of Human Factors in Computing Systems CHI'94* (pp. 445-451), Boston, MA.: ACM Press.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*, 513-541.

Jeffries, R. P., Polson, P. G., Razran, L. & Atwood, M. (1977). A process model for missionaries-cannibals and other river-crossing problems. *Cognitive Psychology, 9*, 412-440.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*, 492-527.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the power law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition*. Hillsdale, New Jersey: Erlbaum.

Payne, S. J., Richardson, J. & Howes, A. (2000). Strategic use of familiarity in display-based problem solving. *Journal of Experimental Psychology-Learning Memory and Cognition, 2*, 1685-1701.

Richardson, J., Howes, A., & Payne, S.J. (1998) A cognitive model of the use of familiarity in the acquisition of interactive search skill. In M.A. Gernsbacher & S. J. Derry (Eds.), Proceedings of the 20th Annual Conference of the Cognitive Science Society (p. 1258). Mahwah, NJ: Erlbaum.

Roberts, S., Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*, 358-367.

Conceptual Combination as Theory Formation

Dietmar Janetzko
dietmar@cognition.iig.uni-freiburg.de
Institute of Computer Science and Social Research
Center of Cognitive Science
University of Freiburg i. Br.
D-79098 Freiburg

Abstract

Conceptual combination is an instance of synthetic problem solving comparable to design or planning. This work reviews evidence supporting the view that the result of such a synthesis has much in common with theory formation. Similar to theory formation inference in conceptual combination can be modeled by using abduction as the principal mechanism to generate hypotheses. However, abduction in itself provides no answer to questions regarding the explanation and selection of hypotheses. Results of two experiments address these issues and provide converging evidence to the view that conceptual combination is a form of theory formation. The results are interpreted within a framework of constraint satisfaction which is assumed to take place on a micro-level (compounding relations) and on a macro-level (principle of parsimony).

Introduction

What is the "glue" between words like *house* and *boat* that allows us to make sense out of the resulting conceptual combination *house boat*? There is a number of indications suggesting that conceptual combination is an instance of synthetic problem solving (like design or planning) that may be qualified as theory formation *en miniature*: First, most of the researchers in the field agree that conceptual combination can be best described in terms of knowledge structures, viz., two or more concepts that are reconstructed locally once they are involved in conceptual combination. Simple concepts in itself are often viewed as condensed theories (e.g., Murphy & Medin, 1985). Second, in many cases conceptual combinations can be paraphrased by a relative clause (e.g., *a house boat is a boat that ...*). In this way it becomes evident that conceptual combinations may add a more fine-grained conceptual schema to the conceptual classification system we already have. This is an example of *taxonomy revision* that is often related to theory formation (e.g., Shrager & Langley, 1990). Third, conceptual combinations provide support in description, explanation and prediction of phenomena, all of which are often taken to be the defining functions of theories (e.g., Brown & Ghiselli, 1955). The descriptive function is perhaps the most obvious one since conceptual combination is one of the major linguistic mechanisms for word formation (Olsen, 2000). The explanatory function is actually at the heart of conceptual combination: whenever we are confronted with a combination of concepts, we cannot but start searching for a coherent explanation that integrates the concepts usually by relation linking or carry over of an attribute. Finally, the predictive function is supported by the fact that selective inheritance of attributes is made possible in conceptual combination (Hampton, 1987). Hence, new phenomena or artifacts are often labeled by making use of it (Costello & Keane, 2000). While work in theory formation has traditionally focused on respectable scientific theories that have become hallmarks in the history of science, theory formation in conceptual combination is of a more mundane type. Usually, its basic function is to set up micro theories that help explaining simple compounds like, e.g., *turpentine jar* and the phenomena they are referring to. However, there are striking parallels between both types of theory formation.

Though most of the work on conceptual combination relies on comparable schemas of knowledge representation, the procedural assumptions may differ considerably. Upon closer inspection it becomes evident that some of the variance in the field is due to the fact that different problems of conceptual combination are focused. These aspects can be sorted by recasting them in terms of a model of theory formation. In so doing, constraint satisfaction needs to be recognized as a major aspect of processing in synthetic problem solving (Smith & Brown, 1993). As in many instances of synthetic problem solving, in conceptual combination there is a huge number of possibilities to integrate entities (e.g., nouns). This is evidenced by the high number of interpretations obtained especially from novel compounds (Costello & Keane, 1997).

However, many of the investigations of conceptual combination rely exclusively on interpretation and rating tasks. These methods can only tap time-consuming processes. Clearly, we all know that conceptual combination can proceed both slowly and controlled. But there is also evidence from the few reaction time studies on conceptual combination that this process can also be very fast and and carried out automatically (e.g., Gagné & Shoben, 1997). The fact that processing of conceptual combinations may be accomplished either very fast or slowly needs to be accounted for. For this reason, I will suggest a schema that addresses the issue of constraint satisfaction in conceptual combination on two lev-

els, which classify problems of conceptual combination according to two tasks:

I. The interface-selection task. A open question in conceptual combination is whether or not conceptual combinations are represented as a whole or in parts (*full listing hypothesis* vs. *decomposition hypothesis*, Butterworth, 1983). This issue clearly has consequences for models of processing of compounds and thus for the interface-selection task in conceptual combination. Usually, however, in work on conceptual combination it is the *decomposition hypothesis*, which is implicitly adopted. In so doing, a number of rationales has been hypothesized for the selection of the part(s) of the knowledge structures involved (e.g., slots, relations) that are taken to establish the linkage between combining concepts. Investigations and models that address the interface-selection task have been put forth: Wisniewski (2000) suggested that an alignment process, viz., similarity assessment, between modifier and head guides this task, while Estes & Glucksberg (2000) found evidence that salience of attributes is underlying interface-selection. Finally, investigations by Gagné & Shoben (1997) supported the view that there is a fixed set of compounding relations and selection of a relation is done according to the frequency of its usage. Construction of interpretations of conceptual combinations proceeds by rendering the reconstructed knowledge of the concepts involved into the natural language. This task is accomplished if one or many interpretations of candidate conceptual combinatations are found.

II. The interpretation-selection task. Even if the selection problem is solved successfully, there is usually a great number of possible interpretations remaining. Hence, a second step has to take over that consists of evaluating the candidate interpretations. Work along these lines has been carried out by Costello & Keane (2000) who provide empirical evidence that search of an appropriate interpretation is narrowed down by the constraints of diagnosticity, plausibility and informativeness.

Basically, the two task of this schema boil down to a generate-and-test approach, which is adjusted to issues of conceptual combination. While there are a number of investigations that address either the first or the second task, there is no organizing framework that integrates work in the field and provides empirical evidence supporting this framework. Only in a few investigations the generative nature of conceptual combination that leads to synthesis of knowledge structures has been spelled out in a sound way. The goal of this paper is to identify and investigate mechanisms of theory formation in conceptual combination. In so doing the schema outlined above provides some methodological assistance by guiding the investigations to both a micro-level and a macro-level.

The paper is organized as follows: First, I am describing the type of conceptual combination that has been investigated in this work. Second, the role of theories in conceptual combination is discussed. Third, following the work of Stickl (1989) and Hobbs, Stickel, Appelt & Martin (1990), I give an outline of abduction, which is assumed to be the generative mechanism that drives theory formation in conceptual combination. While the work on abduction mentioned provides a sound foundation of processing conceptual combinations, there are some empirical questions relating to the generation and selection of hypotheses that are not addressed by this approach. Fourth, according to the two-step schema introduced above, I am presenting two experiments on these issues. The first experiment highlights mechanisms of linking modifier and head in conceptual combination. The structure of preferred theories is investigated in the second experiment. The final discussion places the results into the framework presented initially and considers open questions related to theory formation in conceptual combination.

Conceptual Combinations

The first part of a nominal or noun-noun compound is usually called the *modifier* and the second part is referred to as the *head*. There are various schemata for classifying interpretations of this type of compound, but the one most widely accepted schema seems to be the one introduced by Wisniewski (e.g., Wisniewski, 2000). He distinguishes three types of interpretations: In *relation-linking* interpretations, people explicitly use a relation to explain a compound (e.g., robin snake = a snake that eats robins). *Property* interpretations involve one or a few properties of the *modifier* that are applied to the *head* (e.g., robin snake = a snake that has a red underbelly). *Hybrid interpretations* are not precisely characterized since this category might apply to a conjunction of the constituents or a cross between them (robin canary = a bird that is half canary and half robin). The work on conceptual combination described in this paper is focusing on noun-noun compounds that have a relational interpretation.

The Role of Theories

The seminal paper of Murphy & Medin (1985) is often taken to be the beginning of a line of research that views concepts as condensed theories. In many investigations of conceptual combination that follow this approach, knowledge or theories have not been described very precisely. Still, ample evidence has been collected that background or domain knowledge feeds into conceptual combination (e.g., Hampton, 1997).

In the work presented here, theories may be defined on two levels: On a functional level, theories are conceived as knowledge structures subjects may use for description, explanation and prediction of phenomena of interest. On a representational level, I am adopting a schema for describing both the nominal compounds, background knowledge and for deriving thematic or compounding relations that has been suggested by Hobbs et al. (1990, p. 24f):

$$(\exists x, y)\, turpentine(y) \wedge jar(x) \wedge nn(y, x)$$

The three propositions of this logical form are meant to signify a juxtaposition of two nominals, and *nn* is a placeholder for the compounding relation to be found. The background theory might take the following logical form

$$(\forall y) liquid(y) \wedge etc_1(y) \supset turpentine(y)$$

which denotes that being liquid is among other attributes a feature of turpentine and

$$(\forall e_1, x, y) function(e_1 x) \wedge contain'(e_1, x, y) \wedge liquid(y) \wedge etc_2(e_1, x, y) \supset jar(x)$$

meaning that if the function of something (x) is – among other things – to contain liquid, then it may be a jar.[1]

Abduction in Conceptual Combination

The view on conceptual combination outlined in this work follows a rationale of inference called *abduction* that can be described as explanatory hypothesis generation (Stickel, 1989; Hobbs et al, 1990). More precisely, the proper place of (the generative part of) abduction in the schema described above is within the first task: By abductive inference hypotheses are generated on the basis of domain or background knowledge that provide a means for interface selection. Evaluation of the hypothesis is part of the second task.

Abduction is a mechanism used frequently in models of theory formation. Contrary to deductive reasoning there is no guarantee for correctness in abductive reasoning. Cast in a more concise formal lingo, abduction can be described as follows:

$\mathfrak{D}$ is a collection of data (facts, observations, givens);
(1.) $\mathcal{H}$ explains $\mathfrak{D}$ ($\mathcal{H}$ would, if true, imply $\mathfrak{D}$);
(2.) No other hypothesis explains $\mathfrak{D}$ as well as $\mathcal{H}$;
Therefore, $\mathcal{H}$ is correct.
(cf. Falkenhaimer, 1990, p. 160, numbers added).

Applied to the analysis of conceptual combination $\mathfrak{D}$ refers to a noun-noun juxtaposition. $\mathfrak{D}$ is really just a juxtaposition and does not provide any hints concerning its potential coherence or fitting together. However, it motivates processes that seek to find evidence in favor or against coherence in $\mathfrak{D}$ (cf. Thagard, 1997). $\mathcal{H}$ signifies one or many compounding relation(s). They slip into the role of hypotheses that have the potential of specifying in which way the concepts of $\mathfrak{D}$ cohere. Note that hypotheses are derived from domain or background theories. In our example introduced in the preceding section we may infer abductively

$$(\forall e_1 x, y) contain'(e_1, x, y) \supset nn(x, y)$$

meaning that the placeholder *nn* might be identified with the relation or hypothesis *contains*.

If there are more hypotheses that may explain $\mathfrak{D}$, the best of them is selected. Clearly the criteria of what "best" means in the field of conceptual combination are not specified in this fairly general definition.

Two things should be noted in the definition of abduction as provided by Falkenhainer (1990): First, abduction is a two-step process that bears strong commonalities to the two tasks in conceptual combination described above. Second, to find out whether or not the definition – and thus abduction – holds in conceptual combination, this definition needs to be applied to the field and also further specified. But what does "true" (1.) and what does "well" mean" (2.) in the definition above?

Both issues are essentially empirical questions. Concerning the first of them I am making the conjecture that the hypothesis $\mathcal{H}$ is said to be true iff the relation can be successfully instantiated by the concepts of $\mathfrak{D}$. Whether and to which degree instantiation is modified by similarity on the level of attributes is also an open issue. Concerning the second issue I assume that a hypothesis is said to explain $\mathfrak{D}$ well if it is parsimonious and sound, which is equivalent to the heuristic of Occam's razor.

Experiment 1: Constraints on the Micro-Level

Experiment 1 examines the effect of activation of thematic relations on the process of conceptual combination. Patterns of the compounding relations were varied as the independent variable. This variable was chosen for two reasons: First, by using this variable it could be investigated whether the *full listing hypothesis* or the *decomposition hypothesis* holds in conceptual combination. Second, by using this variable groups of items could be set up that differed in the degree of similarity. Hence, a comparison of different accounts to conceptual combination could be carried out. These are approaches that rely primarily on similarity (e.g., Wisniewski, 2000) vs. approaches in which compounding relations hold the key to conceptual combination (e.g., Gagné & Shoben, 1997). With regard to the schema introduced above, experiment 1 addressed the interface-selection task, viz., the mechanisms leading to the linkage between modifier an head.

Method

A semantic decision task was used to assess conceptual combination in nominal compounds. Subjects were instructed to read both prime and target and were requested to decide as quickly and as correctly as possible whether the target was a concept that refers to a material entity (e.g., rubberball).

Participants. The subjects were 39 students (18 male and 21 female) of Freiburg University who either participated for course credit or payment. The age of the subjects ranged between 18 and 29.

Materials and Procedure. The experimental stimuli were prime-target pairs. Both prime and target were common compounds that were based on simple German

[1] The primed predicate $contain'(e_1, x, y)$ together with its arguments signify that e_1 is the eventuality of *contain* being true for x and y.

nouns[2]. Compounds based on metaphors, names or associated words were excluded. Since novel compounds are known to elicit a variety of interpretations, I used common compounds that have a standard interpretation. In this way, fixation of the number of interpretations was achieved, and thus the effect of the independent variable, viz., the pattern of the compounding relation, could be investigated more precisely. Investigations of the effects of the independent variable led to the selection of 4 groups of items (cf. Table 1).

	CoI	CI	NCI	DI
Prime	lipstick	snowball	tennisball	summertime
Target	rubberball	rubberball	rubberball	speedlimit

Table 1: Groups of Items used in Experiment 1

18 *Control items (CoI)*, which were made of pairs of compounds each of which used a different thematic relation. Moreover, the words in each pair were different (e.g., tennisball - snowball - thematic relations: "x is made of y", "x is used for y"). All items used (both prime and target) were made up of concrete concepts (e.g., lipstick).
18 *Concordant items (CI)*, which were made of pairs of compounds both of which shared the same thematic relation. Moreover, the head concept was identical in prime and target (e.g., snowball - rubberball, common thematic relation: "x is made of y"). To achieve a balance between concrete and abstract relations, 9 of the items of CI used the relation "x is made of y", and 9 employed the relation "x is used for y". All CI used concepts (both prime and target) referring to concrete words.
18 *Non-concordant items (NCI)*, which were made of pairs of compounds each of which used a different thematic relation. Still, 9 target compounds used a concrete thematic relation ("x is part of y"), while 9 used an abstract thematic relation ("x is used for y"). The head concept was identical in prime and target (e.g., tennisball - snowball, thematic relations: "x is made of y", "x is used for y"). All NCI used concepts (both prime and target) referring to material entities.
18 *Distractor items (DI)*, which were made of pairs of compounds each of which used a different thematic relation. In contrast to items of all other groups all distractor items (both prime and target) used abstract concepts (e.g., speed record).

Note that prime and target of CI, NCI and CoI (each of which was presented together with the items from DI) were becoming increasingly dissimilar: In CI there was an identity of thematic relation and modifier, in NCI only the heads were overlapping. Finally, in CoI there was neither on the level of words nor on the level of the thematic or compounding relations any overlap. While all words in CI, NCI and CoI were concrete, all words in DI were abstract.

[2]Note that the experiment was carried out in German where all compounds are written as one word.

The items were used in a between subjects design with three groups. Subjects of each group was presented with the 36 Items of (CI & DI; NCI & DI, CoI & DI). Thus, in each group there was the same number of abstract and concrete targets (18:18). The SOA was 300 msec and the ISI was 100 msec (cf. Zwitserlood, 1994). Subjects worked first through a series of 24 training items that included a mixture of all types of items mentioned above. After that subjects were requested to decide as quickly and as correctly as possible whether the 36 items used were concrete or abstract words.

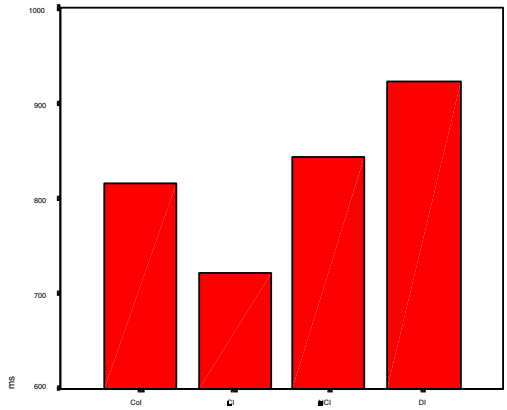


Figure 1: Results of Experiment 1

Results and Discussion

Fig. 1 presents an overall view of the results of experiment 1. Pairwise analyses of the results were carried out via Mann-Whitney-U-tests and via the Wilcoxon-test in the case of the dependent samples involved in the comparison of CI and DI. Scores of CI were significantly lower than the scores of the groups NCI (z=-7,37 , $p \leq .001$), CoI (z=-3,41 , $p \leq .001$), and DI (z=-6,45 , $p \leq .001$). Interestingly, the difference between scores of CoI and scores of NCI was not statistically reliable. Beyond that the difference does not point into the direction expected on the basis of a similarity approach (cf. Figure 1). This suggests that similarity (on the level of attributes) between heads of prime and target had no facilitating effect.

The results indicate two things: First, processing of the compounding thematic relation plays indeed an important role in conceptual combination. Throughout the investigation the thematic relation has never been expressed explicity. Given the fast mode of the task we may safely conclude that the compounding relation is processed unconsciously. This is especially striking since common compounds were used. These concepts are often believed to be processed as one unit without considering the constituents. This can be taken as a conservative test of the *decomposition hypothesis*, which was clearly better supported by the data than the *full listing hypothesis*.

Second, having addressed the more basic question whether the *decomposition hypothesis* or *full listing hy-*

pothesis gives a better account of the data, I will now turn to the question whether similarity (on the level of attributes) or relations hold the key for the interface-selection task. It is worth pointing out that NCI scored quite low although there was an identity of heads in both prime and target. If attributes had played at least a minor role, then the increasing similarity on the side of the stimuli (*sim* CI > NCI > CoI) would have induced corresponding effects on the side of the dependent variable (*rt* CI > NCI > CoI). However, this is not the case. Taken together, the data do not support the view that similarity (as specified on the level of attributes) provides the rationale of addressing the selection problem.

It is tempting to assume that by priming a particular (misfitting) relation in NCI, this relation may block or reduce the salience of the most suitable thematic roles of the constituents. Hence, the subject has to make an effort to retrieve a more appropriate relation from the domain knowledge. This may be due to a time-consuming derivation process. A possible model of this process is provided by the abductive rationale spelled out by Hobbs et al. (1990).

Experiment 2: Constraints on the Macro-Level

The goal of experiment 2 was to examine aspects of compound interpretation that affect its acceptance. With respect to the schema introduced above, experiment 2 addressed the interpretation selection task, viz., the choice between competing interpretations of a conceptual combination. If conceptual combination is indeed a form of theory formation, then features like "concise" and "plausible" often considered to be aspects of a good theory should also characterize an appropriate interpretation of a conceptual combination.

Method

In experiment 2 judgments of interpretations of novel compounds were elicited. The type of interpretation was used as the independent variable. The interpretations employed in the experiment had been generated and assessed in two preparation studies conducted before experiment 2 with independent samples of subjects.

Participants. 121 subjects (57 male, 64 female) between 16 and 42 years old participated in experiment 2.

Materials and Procedure. 4 × 20 pairs of novel compounds along with group specific interpretations were used in experiment 2. The material consisted of novel compounds since common compounds have a standard interpretation. Thus, a variety of different possible and in principle equally appropriate interpretations could not be generated on the basis of common compounds.

The interpretations investigated in experiment 2 had been set up in two preparation studies carried out with different subjects: First, in an in-between study a sample of subjects (20 subjects, 12 female, 8 male, between 18 and 41 years old) was requested to generate interpretations of 20 novel German compounds (e.g., "curtain hotel") according to 4 conditions: *detailed and creative* (dc), *detailed and plausible* (dp), *concise and creative* (cc), and *concise and plausible* (cp). Second, in a subsequent in-between rating study a new sample of subjects (32 subjects, 21 female, 11 male, between 18 and 47 years old) assessed each group of interpretations on a five-point scale according to its aptness or inaptness. For each of the 20 novel compounds that have been used, 4 interpretations with the highest aptness ratings were selected and were subsequently employed in experiment 2.

The final outcome of the two preparation studies were 80 pairs of conceptual combination + interpretation all of which were on a high level of aptness. The pool of 80 pairs of conceptual combination + interpretation was divided into 4 groups of 20 items. Each group consisted of an equal share of items from all 4 conditions (dc, dp, cc, cp). In experiment 2, subjects of each group were presented with 20 pairs of novel compounds + interpretations that should be assessed on a five point rating scale (1 excellent – 5 unappropriate) concerning their aptness as an explanation of the compound.

Results and Discussion

Fig. 2 presents an overall view of the results of experiment 2. Pairwise comparisons of the results were conducted via Mann-Whitney-U-Tests. Scores of cp turned out to be significantly lower than the score of the groups dc (z=-16,29 , p <.001), dp (z=-3,54 , p <.001), and cc (z=-18,05 , p <.001).

The results of experiment 2 show that criteria that are usually applied to sound theories also apply to interpretations of conceptual combinations. These results are consistent with the more general hypothesis of this paper that conceptual combination is a form of theory formation.

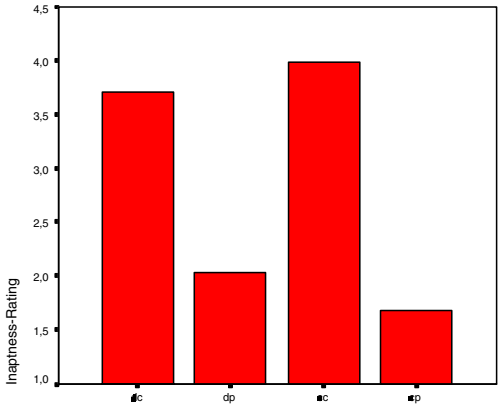


Figure 2: Results of Experiment 2

General Discussion

Research on conceptual combination can be characterized by two more general issues: First, empirical work in the field has mostly been conducted isolated from considerable formal work on conceptual combination. The

formal work on abduction outlined briefly in this paper provides valuable insights for conceptual combination, e.g., concerning knowledge representation and abduction. Second, within the camp of empirical researchers the two basic tasks in conceptual combination have usually not been distinguished properly. The work presented in this paper reacts to these problems both by setting up an account of conceptual combination as theory formation that integrates many aspects of the field and by providing empirical data that is consistent with this framework. Clearly, more empirical work is necessary that fleshes out the framework presented in this work and elucidate the role of domain or background knowledge which has also been found influencial in conceptual combination. In fact, work on abduction coming mostly from computational linguistics offers some guidance for knowledge representation (Hobbs et al., 1990, p. 24), which is almost absent in more psychological work on this topic. Taken together, both issues stress the role of abduction in conceptual combination. This pattern of reasoning might help to explain conceptual combination in terms of theory formation.

Viewing conceptual combination as an example of theory formation holds the promise of a crossfertilisation between two hitherto almost uncombined research tradition: Research on conceptual combination could became more aware than hitherto that the phenomenon under study is a generative process details of which can be captured in terms of explicit schemas of knowledge representation. On the other hand, work in theory formation that has been focusing on theory formation in the natural science (an excellent survey is given by Darden, 1997) could broaden this perspective and consider theory formation *en miniature* in conceptual combination. Investigations based on this research strategy could be fruitfully applied in psychology, anthropology and ethnology.

Acknowledgments

I like to thank Zachary Estes, Barbara Hemforth, Bipin Indurkhya, and Gerhard Strube for valuable comments to this paper, and I am grateful to Roman Kennke for the programming work put into this research project.

References

Brown, J. S. & Ghiselli, E. E. (1955). *Scientific method in psychology*. McGraw-Hill, New York.

Butterworth, B. (1983). *Language processes (Vol 2)*. Academic Press, London.

Costello, F. & Keane, M. T. (1997). Polysemy in conceptual combination: Testing the constraint theory of combination. *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 137–141). Hillsdale, NJ: Lawrence Erlbaum Associates.

Costello, F., & Keane, M.T. (2000). Efficient Creativity: Constraints on conceptual combination. *Cognitive Science*, *24*, 299-349.

Darden, L. (1997). Recent work in computational scientific discovery. *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 161–166). Hillsdale, NJ: Lawrence Erlbaum Associates.

Estes, Z. & Glucksberg, S. (2000). Interactive property attribution in concept combination. *Memory & Cognition*, *28*, 28-34.

Gagné, C. L. & Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 71–87.

Hampton, G. (1987). Inheritance of attributes in natural concept conjunctions. *Memory and Cognition*, *15*, 55–71.

Hobbs, J., Stickel, M., Appelt, D., and Martin, P. (1990). Interpretation as abduction. *Technical note 499, Menlo Park, CASRI International.*

Murphy, G. & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289-316.

Shrager, J. & Langley, P. (1990). Computational approaches to scientific discovery. In J. Shrager & P. Langley (Eds.), *Computational models of scientific discovery and theory formation.* (pp. 1-25). San Mateo, CA, Morgan Kaufmann.

Smith, G. F. & Brown, G. J. (1993). Conceptual foundations of design problem solving. *IEEE Transactions on Systems, Man, and Cybernetics*, *23*, 1209-1219.

Stickel, M. (1989). Rationale and Methods for Abductive Reasoning in Natural Language Interpretation. In: R. Studer (Ed.), Natural Language and Logic, Berlin: Springer, 233-252.

Thagard, P. (1997). Coherent and creative conceptual Combinations. In T.B. Ward, S.M. Smith, & J. Viad (Eds.), *Creative thought: An investigation of conceptual structures and processes.* (pp. 129-141). Washington, D.C.: American Psychological Association.

Wisniewski, E. (2000). Similarity, alignment, and conceptual combination: Comments on Estes and Glucksberg. *Memory & Cognition*, *28*, 35-38.

Zwitserlood, P. (1994). The role of transparency in the processing and representation of dutch compounds. *Language and Cognitive Processes*, *9*, 341-368.

Combining Integral and Separable Subspaces

Mikael Johannesson (mikael.johannesson@ida.his.se)
Department of Computer Science, University of Skövde, Sweden and
Lund University Cognitive Science, Lund University, Lund, Sweden

Abstract

It is well known that pairs of dimensions that are processed holistically - *integral dimensions* - normally combine with a Euclidean metric, whereas pairs of dimensions that are processed analytically - *separable dimensions* - most often combine with a city-block metric. This paper extends earlier research regarding information integration in that it deals with complex stimuli consisting of both dimensional pairs previously identified as holistic, and dimensional pairs previously identified as analytical. The general pattern identified is that information integration can be more accurately described with a rule taking aspects of stimuli into consideration compared to a traditional rule. For example, it appears that combinations of analytical and holistic stimuli, are better described by treating the different subspaces individually and then combining these with addition, compared to any single Minkowskian rule, and much better compared to any of the Minkowskian rules traditionally used (i.e. the city-block-, the Euclidean or the dominance-metrics).

Introduction

Suppose we have objects that differ on several dimensions – how is (dis-) similarity of such objects related to (dis-) similarity on each of the dimensions? Since Attneave (1950) raised essentially this question, much research efforts have been focused on the applicability of different combination rules. The most commonly investigated rules, or metrics, for describing distances in a multidimensional space have been instances of the generalised Minkowski metric (Eq. 1).

$$(1)\quad d(i,j) = \left\{ \sum_{k=1}^{n} | x_{ik} - x_{jk} |^{r} \right\}^{1/r} \quad ; \; r \geq 1$$

where $d(i,j)$ is the distance between object i and j, x_{ik} refers to the position of object i on the kth axis and n is the number of constituting dimensions.

Three extreme cases can be identified: r = 1; *the city-block metric* - The distance is the sum of the absolute differences for each of the underlying dimensions; r = 2: *the Euclidean metric* - The distance corresponds to the square root of the sum of the squared differences for each of the underlying dimensions; and r = ∞ : *the dominance metric* - The distance between two objects is a function of the distance for the separate dimension that differ the most.

When it comes to cognitive modelling, the city-block and, especially, the Euclidean metrics are the most common. However, it is well established that some pairs of dimensions combine, with Garners (1974) terminology, to form *integral dimensions* and others to form *separable dimensions* (see e.g. Garner, 1974, 1977). An integral pair typically is processed as holistic, unanalysable, directly and effortlessly by subjects and that the constituent dimensions combine so as to conform to a Euclidean metric; pairs of hue, saturation or brightness of colour (see e.g. Hyman & Well, 1967; Kemler Nelson, 1993) and the auditory dimensions of pitch and loudness (Kemler Nelson, 1993) typically do this. The corresponding description for a separable pair is that the constituent dimensions are processed independently by subjects and that they combine so as to conform to a city-block metric, e.g. size and reflectance of squares (Attneave, 1950).

Now, the combination rules for integral and separable dimensions are well investigated for dimensional *pairs*. But, what about more complex combinations? How do we integrate information when both integral and separable pairs are involved? Adequate descriptions of information integration behaviour is not only important from a theoretical perspective, but also from a more practical and pragmatic machine learning perspective.

Simple parallelograms varying in saturation, brightness, height and tilt could serve as an example. Pairs of the dimensions of colour, i.e. of hue, brightness and saturation, are often used as prototypical examples of integral dimensions (see e.g. Hyman & Well, 1967; Kemler Nelson, 1993). Perception of variation in saturation and brightness on a single colour patch have in previous studies (e.g. Hyman & Well, 1967; 1968) been shown to be better described using the Euclidean compared to the city-block metric. The height- (size-) and tilt- dimensions of parallelograms is an example of separable dimensions (Tversky & Gati, 1982). Tversky and Gati found such pairs to be better described using the city-block metric compared to the Euclidean.

How, then, could subjects' phenomenological (dis-) similarity between parallelograms varying in height, tilt, saturation and brightness be described? With reference to the different metric properties of the underlying pairs of dimensions, it makes sense to divide the stimuli space into two separate subspaces - one describing the aspects of shape of the stimuli (i.e. height and tilt) - *the shape space* - and one the colour aspects (i.e. saturation and brightness) - *the colour space*. In this case it could be that two different metrics should be

applied: the city-block metric for the shape space and the Euclidean metric for the colour space. For combining the separate subspaces into a holistic measure, simple addition could be expected, with reference to that the pair of subspaces better fit the description of separability compared to integrality.

In the remainder of this paper, combination rules, such that the same Minkowski-r (r) applies to the whole stimuli space, will be referred to as *homogenous* rules. Metrics such that one r, say r_1, applies to one subspace, and one r, say r_2, applies to another, and that the holistic measure is obtained by combining the sub-metrics separably, will in the following be referred to as *heterogeneous* rules or metrics[1].

Now, how could we determine which of reasonable alternatives is the best when we want to describe (dis-) similarity judgements of stimuli varying in height, tilt, saturation and brightness?

General Method

The method used by Dunn (1983) in order to investigate the relationship between dimensional integrality and the combination rule used in a dissimilarity judgement task, will be adopted in this paper. However, it will be generalised in order to deal with stimuli with more than two underlying dimensions.

The basic idea is to divide the set of dissimilarity ratings into unidimensional and bidimensional ratings, reduce them to distances between points in a predefined dimensional space and then determine the r that best predicts the bidimensional dissimilarities from the unidimensional ones. In order to reduce ratings to distances correspondence, interdimensional additivity, intradimensional subtractivity and linearity must be assumed (Dunn, 1983; see also Johannesson, 2001a).

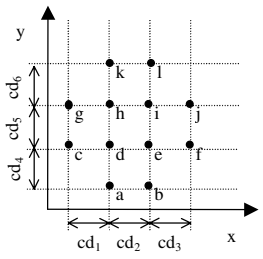


Figure 1: 12 stimuli and their component distances (cd1-cd6). (Based on Figure 1. in Dunn, 1983).

A first step is to decompose the unidimensional distances into component distances (see Figure 1 above). Further, under the assumption that the function relating dissimilarities to distances is linear, the dissimilarities between the stimuli in Figure 1 could, according to Dunn (1983), be expressed as

$$(2)\quad \delta(a,b) = \sum_{i=1}^{6} w_{iab} cd_i + A \;;w_{iab} = 0 \text{ or } 1$$

where $\delta(a,b)$ is the perceived dissimilarity between object a and b, w_{iab} refers to the weight of the component distance cd_i, and A is an additive constant.

Eq. 2 specifies a multiple regression equation in which the weights define a set of dummy variables, the component distances form the regression coefficients and A is the additive constant.

Determining the Spatial Metric

Performing a multiple regression analysis on unidimensional dissimilarities provides an estimate of the component distances and the additive constant. From these, it is straightforward to estimate any "Minkowski dissimilarity". In order to determine the "best" describing metric for a particular subject, Dunn (1983) compared the mean observed and the mean predicted bidimensional dissimilarity using a certain value of r: overestimation of r lead to underestimation of the observed mean, whereas underestimation of r lead to overestimation of the observed mean.

Methodology Adopted

The methodology outlined by Dunn will be adopted with the exceptions as outlined below. Since the present paper aims to investigate whether the machine learning community could gain from using different Minkowski metrics for different subspaces rather than a single metric applied to the whole space, the various tests suggested by Dunn are not central here.

Experiment I

Subjects

14 students at the University of Skövde participated for a reward of cinema tickets roughly worth £11 or $17.

Stimuli

The stimuli were parallelograms varying in height (h; 4, 5 or 6 units of length), tilt (t; 40, 50 or 60 degrees), saturation (s; 40, 60 or 80% of maximum saturation) and brightness (b; 40, 60 or 80% of maximum brightness). The width and the hue of the parallelograms were held constant (4 units of length and 240 degrees of the colour circle, respectively).

In order to not exhaust the subjects, 20% of the possible 3240 pairs (i.e. 648) where chosen randomly. The order of the selected pairs were randomised.

[1] There are few examples of true integral dimensions in the literature (Grau & Kemler Nelson, 1988). This fact does not undermine the possible practical importance of heterogenous models, since perception of many dimensional pairs fall between the endpoints of a continuum of dimensional separability (Smith & Kilroy, 1979; Smith, 1980).

Procedure

The experimental sessions were performed individually in a quiet room with drawn curtains.

Each subject was first asked whether she/he had normal colour vision or not[2], and was then asked to follow the instructions given on the screen.

The experiment consisted of several phases:

- *Instruction phase*: Subjects were informed that they should judge dissimilarity between coloured parallelograms using a 20-graded scale.
- *Stimulus presentation*: Diminished versions of all stimuli were presented simultaneously in a randomised layout.
- *Training phase*: Subjects made dissimilarity judgements for ten pairs of coloured parallelograms varying in the same dimensions as the real stimulus material. The levels did not coincide with the levels of the real material.
- *Instruction phase*: An instruction phase as above was repeated. This time subjects were also informed that the judgement sessions would be divided into six parts with breaks between.
- *Stimulus presentation*: Subjects were again presented with the complete stimulus material.
- *Judgement phase*: The 648 stimulus pairs were presented in the same random order for all subjects. The experiment took about 2 hours.

All subjects reported they had normal colour vision.

Results and analysis

Table 1 below presents the average component distances (see Figure 1 above) per dimension, and the coefficient of determination for the collapsed data.

Table 1: Average component distances and R^2.

Avg_h	Avg_t	Avg_s	Avg_b	R^2
4.160	2.907	1.214	1.214	0.762

The average component distances, which could be interpreted as the relative saliency of each dimension (Dunn, 1983), differ between dimensions. Especially, the saturation and brightness dimensions have somewhat shorter component distances (are less weighted) compared to height and tilt. A possible explanation for this unequal weighting is that subjects perceived the variation in height and tilt as larger compared to the variation for the integral pair of saturation and brightness.

The coefficient of determination is not very large, indicating that a linear model misses to account for a considerable proportion of the variance of the data.

Determining the Spatial Metric When there are just two underlying dimensions it is obvious that distances should be estimated and evaluated for stimuli differing in two dimensions. However, as the number of underlying dimensions increases, so does the number of possibilities. In the present case, when four underlying dimensions were used, stimuli pairs differing in two or more dimensions were analysed.

Justifying the Measure of Error In order to possibly improve the process of determining the spatial metric, two alternative measures of error for a particular r were contrasted. One was in line with Dunn's method: deviation of the absolute difference between the mean observed dissimilarity and the mean predicted/estimated dissimilarity from the mean observed dissimilarity - in the following referred to as DEV. The other, referred to as the mean squared error (MSE), is defined as

$$(3)\quad MSE = \left(\sum_{\substack{a,b \\ a \neq b}} \left(\delta(a,b) - \tilde{\delta}(a,b) \right)^2 \right) / N$$

where $\delta(a,b)$ is the perceived, $\tilde{\delta}(a,b)$ is the predicted/estimated - dissimilarity between object *a* and *b*, and *N* is the number of stimuli pairs.

For each of the homogenous rules: city-block, Euclidean and dominance, and all non-ordered combinations of heterogeneous rules, where the subspaces where formed by the city-block, Euclidean or dominance metric[3], the distances between all non-ordered combinations of stimuli were calculated from physical descriptions of the stimuli. By regarding the distances as fictive dissimilarities, and by estimating the dissimilarities as described above for different rules, the errors according to DEV and MSE were calculated. The same subset and physical descriptions as used in the present experiment were analysed. Further, the estimated distances were scaled into a discrete scale ranging from 1 to 20. Since the underlying rule was known in each case, the two alternative measures of error could be evaluated against each other.

For the homogeneous models, both DEV and MSE suggested the same - and correct - underlying model. For the heterogeneous models MSE suggested the correct model in all cases. The use of DEV, however, was clearly systematically ambiguous. In all cases when

[2] In a pilot experiment preceeding this subjects performed a colour test in order to find out if they could discriminate between the colours that were to be used. Since all subjects in the pilot experiment reported the colour test to be simple a simple question was judged to be enough.

[3] Note that the heterogenous rule where both subspaces are formed by the city-block metric exactly corresponds to the city-block homogenous rule.

the underlying model could be described *as metric A applies to subspace 1 and metric B applies to subspace 2*, **both** the correct model and the model such that *metric B applies to subspace 1 and metric A applies to subspace 2*, were suggested. The explanation is that that the sum of absolute deviations for the two models necessarily is the same for a balanced set of stimuli.

In summary, based on this analysis, MSE appear to be the better measure for the purposes of this paper.

Spatial Metric Candidates for describing the individual subjects' data were evaluated using MSE as the measure of error. In addition to the rules used when evaluating the two error measures above, i.e.

- the homogenous rules: city-block, Euclidean and dominance - in the following referred to as *Hom cit*, *Hom euc* and *Hom dom*, respectively,
- all non-ordered combinations of heterogeneous rules, where each of the subspaces were formed by the city-block, Euclidean or dominance metric - in the following referred to as *Het citeuc*, *Het citdom*, *Het euccit*, *Hit euceuc*, *Hit eucdom*, *Hit domcit*, *Hit domeuc* and *Hit domdom*, respectively,

errors were calculated for values of Minkowski-r ranging in small discrete steps from r = 1.0 to r = 50.0 applied to the whole stimuli space (the homogenous model giving the lowest error will be referred to as *Hom opt*), the shape subspace and the colour subspace respectively. The heterogeneous model where the separately optimised r for the two-dimensional shape space is applied to "shape" and the separately optimised r for the two-dimensional colour space is applied to "colour", will be referred to as *Het sepHT-sepSB*. Finally, the combination of r:s, one for the shape subspace and one for the colour subspace, when optimised simultaneously with a heterogeneous rule - will be referred to as *Het simHTsimSB*.

Table 2: Models, r:s and errors for average data.

	R	Err
Het simHTsimSB	1.55;2.25	2.146
Het sepHTsepSB	1.55;2.2	2.146
Het euceuc	2;2	2.339
Hom opt	1.2	2.481
Het eucdom	2;50	2.601
Het euccit	2;1	2.894
Het domeuc	50;2	3.838
Het domcit	50;1	3.905
Het citdom	1;50	3.948
Het citeuc	1;2	4.194
Het domdom	50;50	4.313
Hom cit	1	5.907
Hom euc	2	7.805
Hom dom	50	16.644

The candidate models evaluated and the errors for the collapsed data are presented in Table 2 above.

A heterogeneous model combining a rule between the city-block and the Euclidean metrics[4] for the shape space, and a rule roughly corresponding to the Euclidean metric for the colour space (*Het simHTsimSB* and *Het sepHTsepSB*), gave a lower error than the best of the homogenous models (*Hom opt*), which had a r = 1.2, i.e. halfway between the city-block and the Euclidean metrics. This was true irrespectively of if the r:s were optimised separately or simultaneously. The optimal heterogeneous Minkowski-r:s found were lower for the shape space compared to the colour space. However, the r:s found were slightly different from the levels identified by previous research when two-dimensional stimuli have been used. The values were somewhat higher compared to what has been identified for these spaces before. It may be the case that the r-value goes up when the dimensionality increases. This speculation makes sense considering that we have limitations in terms of how many dimensions we can process simultaneously, and that larger values of r corresponds to focusing more on the dimension where the stimuli-pair at hand differ the most.

The common homogenous Euclidean rule (*Hom euc*) gave a substantially worse error than both the best heterogeneous rule and the best homogenous rule. However, the somewhat unequal weightings of the dimensions defining the two subspaces (see above) probably causes the peculiarity that *Het euccit* produces an error lower than that for *Het citeuc*. The fact that there are differences in weighting indicate that there are differences in salience between dimensions.

In summary, a heterogeneous rule or model seems to describe the data better compared to a homogenous one.

Errors and r:s were calculated also for the heterogeneous rules combining the "odd", or counterintuitive, subspaces h/s and t/b on one hand and h/b and t/s on the other. The heterogeneous models with the lowest errors for the average data for each of the three subspace divisions are presented in Table 3.

Table 3: Different subspace divisions and their errors.

Subspace division	Model	Err
height/tilt; sat./bri.	Het simHTsimSB	2.146
height/sat.; tilt/bri.	Het simHSsimTB	3.030
height/bri.; tilt/sat.	Het simHBsimTS	2.861

The errors for the heterogeneous models for the "odd" subspace divisions are considerably larger compared to the error for the original division. For the

[4] Note, however, that the Minkowsky-r of a rule giving distances halfway between the city-block and the Euclidean metric is not the intuitive 1.5, but rather approximately 1.2.

individual data, the corresponding difference was true for 8 out of 12 cases with at least one r differing from 1.0. This difference indicate that the intuitive division into subspaces of shape and colour makes sense.

Experiment II

In Experiment II, the heterogeneous r:s found were larger than what has been found in earlier research. A reasonable question is if the element of non-separability together with the increased dimensionality causes such effects. A second experiment was conducted in order to investigate if integrality (non-separability) could be eliminated as an explanation or not. Contrary to Experiment I, the underlying dimensions in the present experiment are purely separable.

Subjects

12 students (the majority were undergraduates) at the University of Skövde participated for a reward of cinema tickets roughly worth £11 or $17.

Stimuli

The stimuli varied in four dimensions, height (h), tilt (t), width of a stripe parallel to the horizontal axes (st) and brightness (b) of a parallelogram. These dimensions differ from the ones used in Experiment I above in some crucial aspects. One is that they do not form intuitive subspaces. Another is that all possible pairs of dimensions match the description of separable dimensions.

Each dimension varied in three levels, h: (4, 5 or 6 units of length), t: (40, 50 or 60 degrees), st: (1, 2 or 3 units of width) and b: (40, 60 or 80% of maximum brightness). The width, hue and saturation were held constant (4 units of length, 240 degrees and 60% of maximum saturation, respectively).

The same pairs (w.r.t. the numbers of the stimuli), and order between pairs as in Experiment I were used.

Procedure

The experiment was conducted as Experiment I above.

Results

The average component distances for the collapsed data in Experiment II (Table 4 below), are not perfectly equal, especially the brightness dimension is weighted less compared to the others.

Table 4: Average component distances and R^2.

Avg_h	Avg_t	Avg_st	Avg_b	R^2
2.089	2.381	1.530	0.625	0.541

The coefficient of determination is very low, hence a general linear model does not apply well.

Spatial Metric The same candidate models as evaluated in Experiment I were evaluated. The resulting errors for the collapsed data are presented in Table 5.

Table 5: Models, r:s and errors for average data.

	r	Err
Het simHTsimSTB	1.1; 1	3.483
Hom opt	1	3.523
Hom cit	1	3.523
Het sepHTsepSTB	1.6; 1	4.010
Het citeuc	1; 2	4.213
Het euccit	2;1	4.434
Het citdom	1; 50	4.638
Het euceuc	2; 2	5.573
Het domcit	50; 1	5.885
Het eucdom	2; 50	6.185
Het domeuc	50; 2	7.234
Het domdom	50; 50	7.935
Hom euc	2	11.333
Hom dom	50	17.219

It is clear that the best rule, of the ones tested for, for describing the collapsed data in Experiment II is close to a city-block rule (*Het simHTsimSTB* (r=1.1;1), *Hom opt* (r=1) and *Hom cit*). It is not, in this special case, possible to view this as supporting either of homogenous or heterogeneous models since the city-block metric is the sum of the differences for the constituting dimensions. Therefore, there is no difference between a homogenous city-block rule and a heterogeneous rule where city-block rules are used within all subspaces.

As opposed to experiment I, the Minkowski-r values (for the best models) did not increase in magnitude with increased dimensionality.

The heterogeneous models with the lowest errors for the average data for each of the three subspace divisions are presented in Table 6. As, for the collapsed data, the optimal "heterogeneous" rule for the "original" subspace division was close to the city-block metric for both subspaces, this was necessarily the case also for the "odd" subspace divisions.

Table 6: Different subspace divisions and their errors.

Subspace division	Model	Err
height/tilt; str./bri.	Het simHTsimSTB	3.483
height/str.; tilt/bri.	Het simHSTsimTB	3.523
height/bri.; tilt/str.	Het simHBsimTST	3.523

General Discussion

The aim of this paper is to investigate and communicate the idea that division of features/dimensions of objects into separate subspaces - when applicable - possibly could increase descriptive power.

Experiment I involved pairs of dimensions previously found to be combined best by the city-block and the Euclidean metric, respectively. The Euclidean rule turned out to badly describe the data. Instead, a heterogeneous rule combining the two subspaces formed by the intuitive division, was found to provide the best description. The r:s for the two subspaces found in this experiment rhymes with previous research in that they really possess different metric properties and that the r for saturation/brightness was higher than for height/tilt. However, both r:s found were somewhat larger compared previous findings for the separate two-dimensional subspaces. The dimensions involved in Experiment I were all expected to be pairwise separable. Also in the four-dimensional case, the best describing metric turned out to be the city-block rule.

The idea presented received support in that the general pattern identified from the experiments is that phenomenological dissimilarity can be more accurately described with a heterogeneous rule taking aspects of the stimuli into consideration, compared to a homogenous Minkowski-metric.

There are a number of open questions. One relevant issue is how the subspaces themselves should be combined. In this paper, only one of many possible ways was investigated. Another question concerns the magnitudes of the r:s identified. Since the r:s estimated in Experiment II were not larger compared to what could be expected for pairwise combinations of the constituent dimensions, it is apparent that the increase in magnitude of r:s as found in Experiment I, is not generalisable to all complex stimuli. However, it is in the developmental literature well documented that the separability changes with experience (see e.g. Smith, 1980), with the direction from integrality to separability. This pattern also apply to short term learning (Johannesson, 2001b). A possible reason for the relatively large r:s in Experiment I could thus be that stimuli with contents of integrality are harder to "learn" than stimuli composed by separable dimensions. If so, the r:s could possibly stabilise at a lower magnitude for sufficiently experienced subjects. If not, it could simply be that the specific metric properties associated with integral/separable dimensions only are true in the context of single pairs of dimensions, i.e. depending on if they are combined or not. An interesting set of stimuli that could be used in order to explore this (and others) issue further is multimodal stimuli composed of the pairwise integral dimensions of pitch/loudness and hue/saturation.

The results presented clearly motivates further research on the idea that information integration could be described as a combination of distances within different subspaces. More research on if, how and when information integration behaviour can be described in terms of combinations of subspaces may shed light on how we interact with the inherently high-dimensional real world. For example, Edelman & Intrator (1997) discuss the necessity of low dimensionality for learning in perceptual tasks - known as 'the curse of dimensionality'. However, even if we always use low-dimensional representations internally, even for cognition, if these representations involve more than two dimensions, cognitive science have interesting problems to solve.

Acknowledgements

This project have been financed by the Department of Computer Science at University of Skövde and by a grant to the Center of Learning Systems at University of Skövde from the Foundation for Knowledge and Competence Development (1507/97), Sweden. I wish to thank Professor Lars Niklasson, University of Skövde, and Professor Peter Gärdenfors, Lund University Cognitive Science, for support and comments.

References

Attneave, F. (1950). Dimensions of similarity. *American Journal of Psychology, 63*, 516-556.

Dunn, J. C. (1983). Spatial metrics of integral and separable dimensions. *Journal of Experimental Psychology: Human Perception and Performance, 9 (2)*, 242-257.

Edelman, S. and Intrator, N. (1997). Learning as formation of low-dimensional representation spaces. In Shafto, M.G. and Langley, P., editors, *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, Erlbaum, Mahwah, NJ, 199-204.

Garner, W. R. (1974*). The processing of information and structure*, New York, Wiley.

Garner, W. R. (1977). The effect of absolute size on the separability of the dimensions of size and brightness. *Bulletin of the Psychonomic Society, 9 (5)*, 380-382.

Hyman, R. and Well, A. (1967). Judgments of similarity and spatial models. *Perception & Psychophysics, 2 (6)*.

Hyman, R and Well, A. (1968). Perceptual separability and spatial models. *Perception & Psychophysics, 3*, 161- 165.

Johannesson, M. (2001a). The Problem of Combining Integral and Separable Dimensions. *Tech. rep. at the Department of Computer Science*. University of Skövde, Sweden., and *Lund University Cognitive Studies*. Lund University , Sweden.

Johannesson, M. (2001b). Toward Separability During Learning. *Submitted for publication*.

Kemler Nelson, D. G. (1993). Processing integral dimensions: The whole view. *Journal of Experimental Psychology: Human Perception and Performance, 19 (5)*, 1105 - 1113.

Tversky, A. and Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review, 89 (2)*, 123-154.

Distributed Cognition in Apes

Christine M. Johnson (johnson@cogsci.ucsd.edu)
Tasha M. Oswald (tashason@hotmail.com)
Department of Cognitive Science
University of California San Diego
La Jolla, CA 92093-0515 USA

Abstract

Socio-cognitive complexity, as manifest in dynamic multi-party interactions, is central to many theories of hominid cognitive evolution. To assess the rudiments of these abilities in apes, we chose to study the triadic interaction known as "social tool use". In this scenario, one animal (the User) is said to interact with a second (the Tool) to in some way influence the behavior of a third (the Target). The "Machiavellian" model typically used to account for such behavior depends on folk theoretical descriptions of intentional manipulation and deception and has mired the field in unproductive attempts to define such mental states and justify their inference based on observable behavior. As an alternative, we conducted a distributed cognition analysis of these interactions among a group of seven bonobos (*Pan paniscus*) at the San Diego Wild Animal Park. Assuming that attentional behavior was critical, we gained insights from a wide range of research including human developmental, neurological, ethological, and comparative experimental work on social gaze in primates. This work lead us to perform a micro-analysis of 16 video segments (12 of social tools, 4 controls) in which we recorded all changes, at 1/6 second intervals, in relative trajectory, relative body and head orientation (open, peripheral, or closed to other animal) and, whenever possible, gaze. We also recorded the source, timing and duration of social interactions (groom, aggress, etc.) that occurred in those segments. In addition, data on patterns of association and social interaction collected over several years were used to establish long-term relationships such as rank and affiliation to situate and thus help interpret the micro-analysis. To facilitate cross-dimensional comparisons, the results of the micro-analysis were represented along time-lines (X axis = time) such that, for a given dyad and a given attentional dimension, each animal's proximity to physical or eye contact was indicated by the proximity of its line (along the positive or negative Y axes) to zero. Transitional probabilities for a change in any one of Animal A's dimensions being immediately followed by a change in any one of Animal B's dimensions ("triggering"), and for simultaneous changes ("syncs"), were computed for each dyad. Our results indicate that, while overt social behavior was directed by the User to the Tool, of the three dyads involved, the User/Target dyad were disproportionately responsive to one anothers' attentional behavior (while their counterparts in the control segments were not). Thus, this distributed approach has generated evidence that bonobos can and do monitor and respond to the attentional states of others, and that "social tool use" in these animals can best be characterized as a rudimentary capacity for multi-tasking.

Introduction

Studying Socio-Cognitive Complexity

Among the practices that set humans apart from other animals are the complex multi-party interactions in which we frequently engage. Comparative research suggests that the cognitive processes involved in such interactions are unusual in that they enable us to deal with a variety of dynamic social parameters simultaneously (e.g. Humphrey 1976). However, the elaborate collaborations (including conversation and other cultural activities) typical of our species present a daunting challenge to the evolutionary theorist concerned with their origins and course of development. By studying complex, polyadic interactions in our closest genetic relative, the ape, we hope to help identify rudiments shared by the less sophisticated system which, during hominid evolution, may have been subject to additional adaptive specialization.

"Social complexity" has been fairly well defined and documented in nonhuman primates. While in most social animals, power relations can be represented by a simple linear hierarchy and get played out in dyadic interactions, in many higher primates (monkeys and apes), rank and power are not necessarily equivalent (De Waal 1986). That is, for example, through triadic interactions such as coalitions or third party interventions, lower ranking animals can sometimes jointly gain access to resources over higher ranking individuals. Furthermore, the actions taken by a given participant have been shown to take into account not only that individual's own relationships to the others, but the relationship that exists between the others as well (e.g. de Waal & Van Hoof 1981; Cheney, Seyfarth & Silk 1995). For example, a subordinate macaque will

recruit aid against an opponent based more on the ally's rank relative to the opponent than on its own rank relationship with the ally (Silk 1999).

However, the concomitant models of the cognition involved are much less well developed. This is, in part, due to the difficulties inherent to studying social cognition. Gaining experimental control over the relevant variables in a social negotiation, for example, is extremely difficult and the relatively few attempts made with nonhuman primates have produced data that are ambiguous and controversial (see Heyes 1994; Povinelli 1994). Furthermore, the theoretical framework typically used to describe such interactions depends on "folk theoretic" accounts of "Machiavellian intelligence" (Byrne & Whiten 1988). That is, the animals involved are anthropomorphically portrayed as engaging in maneuvers designed to outwit, deceive, or otherwise strategically manipulate one another. This approach raises issues concerning complex, internal mental states such as intentionality, self-consciousness, theory of mind, etc. and has mired the field in attempts to define such states and justify their inference based on observable behavior.

In some respects, given our close genetic relationship, our largely-shared repertoire of gesture and expression, and our many brain similarities, such anthropomorphic descriptions of, especially, ape behavior may not be entirely unjustified. However, they constitute an unsatisfactory model of cognition for several reasons. For one, folk-theoretic accounts of mental behavior are arguably metaphorical, and their validity has been challenged even when applied to humans (Churchland 1981). Second, they may well reveal more about the 'storyteller' than about the players to whom they are applied, however consistent the tales that are generated. Third and most critically, it is unclear whether or not, as hypotheses, they are ultimately falsifiable. In any case, while fermenting much interesting discussion over the past decade or so, this approach has yet to yield much in the way of useful data.

An Alternative Approach

In this paper, we are proposing an alternative approach to the study of the cognitive complexity involved in primate negotiations. The distributed cognition approach (e.g. Lave 1988; Rogoff 1990; Resnick , Levine & Teasely 1991; Fogel 1993; Hutchins 1995) takes cognition as a co-constructed, real-time process that occurs not only whithin but also *between* individuals. (See also Vygotsky 1978; Wertsch 1985.) As a result, the observable group processes that occur during social discourse can themselves be taken as cognitive events. Through a detailed analysis of changes propagated across the "media" (Hutchins 1995) of such events, one can chart information flow, characterize task complexity, document developments in the roles of participants, etc. Thus, while not denying that internal mental states are involved, this model offers the advantage, especially as applied to nonhumans, of focusing on the cognition that is *apparent* in social interaction.

In particular, we have chosen to apply this model to the study of the triadic interaction known as "social tool use" (Kummer 1967; Jolly 1985; Goodall 1986) in the bonobo or pygmy chimpanzee. While no formal definition of "social tool" has been established in the primate literature, it is widely applied and commonly understood to involve one animal (the "User") interacting with a second (the "Tool") to in some way influence the behavior of a third (the "Target"). Such interactions are traditionally divided into five basic classes, all of which were included in our database. These include "recruitment" (User attempts to elicit aid from the Tool against the Target), "agonistic buffering" (User physically uses Tool as a shield or otherwise engages with it to diffuse a threat from the Target), "passport" (User positively interacts with Tool who has a special relationship to the Target, such as mother/offspring, to gain access to the Target), "incitement" (User positively interacts with Tool to incite investment from the Target) and "alibi" (User directs exaggerated attention at Tool to avoid responding to a solicitation from, or to disengage from interacting with, the Target).

While expressly avoiding attributing the implied mental states (manipulation, deceit, jealousy, etc.) to the animals involved, we explicitly acknowledge and even adopt these anthropomorphic terms for two reasons. One is that they are useful as 'shorthand' descriptions that readily communicate the essentials of the behavioral dynamics involved. The other is that it was on such an intuitive, anthropomorphic basis that we originally identified the incidents of social tool use from our videotapes of the bonobos' interactions. One goal of this study, then, is to discover, through a detailed analysis of the tapes, whether those intuitions have any basis in behavioral cues that are meaningful not only to us as observers but that also serve as salient, effective media in the animals' distributed cognition. It is from such an empirical base that we may build a stable comparative model of socio-cognitive complexity.

Methodological Challenges

Given that the distributed approach is still in its infancy, especially as applied to nonhumans (although for discussions see Cousi-Korbel & Fragazsy 1995; Strum, Forster & Hutchins 1997; Johnson, 2001) there are few

established protocols for research design and analysis beyond the essentials demanded by the theory. These essentials include using interactions (as opposed to the actions of individuals) as the units of analysis, charting change over time, and collecting observations at multiple time scales, including the historic and the micro-levels (see Hutchins 1995; Fischer & Granott 1995). The latter is generally done via detailed video analysis of particular interactions. The former is done using macro-level sampling techniques over the long-term, and is meant to provide background information (cultural constraints, social relationships, task exposure, etc.) that can help to "situate" and thus interpret the interactions. Beyond such guidelines, however, work in this emerging field faces several challenges in determining what to score, how to represent those data, and how to generate a meaningful interpretation of them.

Challenge I: What to score

For insights into what might be the most relevant parameters to document in the bonobos' behavior, we turned to the ethological traditions embraced by primatologists, the current experimental work by comparative psychologists, related neurological research, and the developmental study of "theory of mind' issues in human children. Interestingly, these approaches converge on some of the elements that they consider critical. For instance, in the developmental work, an emphasis has been placed on joint visual attention and, in general, on the social manipulation of attention, especially gaze (e.g. Baron-Cohen 1995; Moore & Dunham 1995). This has also become the focus of one of the few productive areas of comparative experimental work on social cognition. That is, for example, several primate species have recently been documented as being capable of "gaze following" (e.g. Anderson et al. 1996; Tomasello Call & Hare 1998; Peignot & Anderson 1999). However, with the notable exception of the chimpanzee (e.g. Povinelli & Eddy 1996; Itakura & Tanaka 1998), most respond to the direction of head orientation and seem incapable, even after extensive training, to respond to the eyes-only as a cue (although see Vick & Anderson 2000). Relevant physiological findings include an increase in cortico-steroids and other orienting responses to direct eye contact (e.g. Keating & Keating 1982; Perrett & Mistlin 1990), as well as cerebral cortex cells in monkeys that are sensitive to the direction of head orientation and/or gaze of another (Perrett et al 1985). Our subject, the bonobo, has not yet been tested on any of the above.

Ethologists, likewise, have long been tuned to the attentional behavior of their subjects (see Smith 1977; Alcock 1978). This can be traced to some simple structure/function relationships. For example, across the phyla, animals with forward-facing eyes (like primates) tend to move in the direction their gaze is oriented, making gaze direction a predictive signal for others. Similarly, body orientation (judged in primates by the orientation of the shoulders and degree of proneness) is also a cue as to one individual's readiness to engage with another. However, these measures are complicated by the fact that orienting toward another individual can also be a function of information-gathering. In fact, this dual role of attentional behavior - both sending and receiving information - is responsible for an inherent ambiguity that can be exploited by the likes of primates.

For example, a monkey that does not look directly at another who is threatening it, and thereby does not 'acknowledge' the threat, can sometimes effectively hold the aggressor at bay. Similarly, a subordinate not looking at a contested resource is much less likely to stimulate an overt conflict over it than another who openly displays its interest. In 1967, Chance described a more general relationship between primates' attentional behavior and rank, with subordinates looking to dominants more often than the reverse. This presumably serves to gather information about likely troop movement or other important group developments. Nevertheless, in most primates a direct stare is a threat and the avoidance of eye contact is common in most species (Kummer 1967).

Subject and Procedures

The subjects of this study were seven bonobos (*Pan paniscus*) living as a coherent social group since 1993 at the San Diego Zoo's Wild Animal Park. The group consists of four mature animals - one adult male, two adult females, and one female who went from adolescent to parturient adult over the course of the study. The group also includes three juveniles born, one to each of the females, in 1991, 1994 and 1996.

In response to the above-described research, we chose to score the following dimensions of the bonobos' activity. Historic data on association patterns (per inter-animal distances) and gross social interactions (such as groom, follow, food-sharing, aggress, etc.) were taken in instantaneous scan samples, at two minute intervals, in twenty minute sessions, for months to years preceding the video segments. These data were used to establish long-term rank and affiliative relationships, as well as to document more short-term developments like disputes or alliances, changes in reproductive status, etc.

Approximately fifty hours of videotape taken of the group since 1993 were reviewed by the authors and twelve incidents of 'social tool use' were identified. (Additional putative examples were also found, but

these did not meet practical criteria such as video quality sufficient to assess gaze, or all three animals being visible for at least 80% of the interaction.) These selections were based on the strength of a combined total of ten years of experience watching these animals. The control segments involved the same configuration of animals, from the same periods of time, engaged in the same gross interactions (e.g. a pair grooming in the vicinity of a third) as the test segments, but which did not, in the authors' estimation, involve social tool use.

For each segment, a User, Target and Tool (or their correspondents in the control segments) were identified and data collected on each dyad (User/Target, User/Tool and Target/Tool). The segments, which varied from 15 to 50 seconds long, were analyzed at intervals of 1/6 of a second (ten frames), which was determined as the shortest interval in which any relevant event (e.g. a glance) could take place, For each interval, the following states were recorded for each individual within each dyad: 1) Inter-animal distance, in fractional body widths, 2) Focus of gaze, whenever it could be determined, 3) Relative body orientation and 4) Relative head orientation, the latter two scored as OPEN, PERIPHERAL, or CLOSED. OPEN was defined as (body or head) oriented directly toward the dyad partner. PERIPHERAL was defined as the partner being positioned from about 20o to about 110o right or left of a sagittal plane between the animal's eyes. (This estimate is based on bonobo visual and cranial anatomy and is assumed to include the area in the animal's peripheral visual field. Ankel-Simons 1983) CLOSED was defined as the partner being behind and thus not within the animal's visual field. For any *change* in state across intervals, the individual responsible (i.e. whose movement resulted in the change) was scored as having shifted TO or FROM its partner. If an individual actively maintained a state (e.g. turned its head to track another's movements, thus actively maintaining an OPEN) this type of 'change' was scored as a KEEP. In an ongoing, time-locked narrative, we also recorded all social interactions (such as groom, play, sex, etc.) and individual tension/relaxation indicators (such as scratch, flinch, clap, playface, etc.).

Challenge II: How to represent

When continuous data are collected over multiple media simultaneously, representing the results in a comprehensible form poses another challenge. In general with such data, some sort of time-line charting the course of changes is commonly employed. But, for example, warranting the ordering of the Y axis values, or aligning the valences of such changes across multiple lines (so that, for instance, movement away from the X axis at time T along one dimension is augmented by movement away from the X axis, at the same time, along another dimension) demand a careful assessment of the media being studied. We again looked to behavioral and physiological data on primates to facilitate this assessment.

In our own earlier work with the bonobos, we found that 70% of all incidents of eye contact were followed by a decrease in inter-animal distance and/or direct interaction (of both positive and negative kinds). Bonobos are also unique among nonhuman primates in that they regularly engage in face-to-face sex, during which eye contact is typically made (Savage-Rumbaugh & Wilkerson 1978). In keeping with the above-described primate research, such data make it clear that eye contact should be treated as central in any representation of the bonobos' gaze relationships. That is, head and gaze states can be organized according to their 'proximity' (i.e. ease of transition) to eye contact. Similarly, body orientations and inter-animal distances can likewise be organized according to the accessibility they offer to (physical) contact.

As a result, we have chosen to represent these data in the following manner. For each medium (gaze, body orientation, etc.), each dyad is represented along a single X axis of Time, marked at 1/6 second intervals. One animal's behavioral values are represented along the positive Y axis and the other's along the negative Y axis. In both cases, the Y values farthest from 'contact' are farthest from the X axis, so that when the pair are far away from each other or oriented away from (CLOSED to) each other, their time-lines are far from the X axis. As each makes a move toward, or turns toward, the other, their lines draw closer, until when contact (physical or eye contact) occurs, the lines meet at the X axis. Using these multiple time-lines, we can then chart a "cognitive trajectory" for each dyad, as informative changes move across and between the media. Plus, since the various time lines are constructed according to similar criteria, we can collapse or blend them into a single (paired) line that depicts all the changes (i.e. all TO's, FROM's and KEEP's) in a given dyad's attentional access to one another. Such summary lines are also useful for examining *inter*-dyad relationships, facilitating the search for the constraints imposed not just by the activity of an individual's partner, but also by the joint activity of the other dyad.

Challenge III: Analysis

While we can reasonably assume that the media we are scoring are perceptually accessible to our subjects, we are primarily concerned with whether these media are *functionally salient*. More specifically, we are questioning whether the three dyads involved

differentially produce and respond to changes in each other's attentional behavior.

To address these questions, we will first compare the transitional probabilities (see Bakeman & Gottman 1997), across dyads, of any change in one partner's media being followed, in the next 1/6 second interval, by any change in the other partner's media. We will also compare the occurrence of partners' simultaneous changes and coordinated active maintenance (i.e. simultaneous KEEPs) across dyads. Thus, both synchrony ("SYNCS")and immediacy of reaction ("TRIGGERS") will be taken as measures of the sensitivity of our animals to the attentional behavior of others. In addition, by assessing patterns of change initiation (e.g. Animal A:TO, followed by Animal B:FROM) we aim to establish a motivational valence for these interactions, based on the basic assumption that organisms approach stimuli they consider desirable and avoid those they consider undesirable. Ethologists have long maintained that "approach/avoidance" conflicts provide a reasonable account of much inter-animal positioning as well as the thresholds for dynamic change. Finally, we will also examine our data for large-scale temporal patterns, both in specific state changes as well as along the more general TO/FROM/KEEP dimension.

Preliminary Results

Defining social tool use as: "One animal using another to manipulate or influence the behavior of a third", we asked two additional observers, familiar with the animals but blind to the cognitive model being tested, to judge whether or not the test and control segments met those criteria. These observers each disagreed only on a single different example. In interviews after the tests, the observers indicated that the User's behavior toward the Tool seemed "insincere", at times because of the abruptness of its termination but most often because of "furtive" looks directed by the User to the Target (although neither mentioned a co-sensitivity to attention between both User and Target - see below).

We are currently conducting the micro-analyses of the video segments. Preliminary results indicate that, in all 12 social tool examples, the User/Target dyad can be identified by their disproportionately high level of responsiveness to changes in one another's attentional behavior. That is, both more Triggering and more Syncs occur in the User/Target dyad than in either the User/Tool or the Target/Tool dyads. In contrast, in the control segments, the pair that was engaged in the primary interaction (i.e. that of the User and Tool in the corresponding social tool examples) show a tendency to be the most responsive to one another's attentional behavior. The additional level of attentional sensitivity between what would be the User and the Target does not appear in the control segments.

As expected, the cognitive trajectories of these interactions frequently jump across media (e.g. an approach by one animal triggers a look-away by another). The most intense (simultaneously multi-media) interactions occur in the User/Target dyad. The directionality of triggers (i.e. who initiates an exchange) often varies between and even within segments. In fact, it may be the case that a tendency to 'vacillate' in the direction of triggers is especially characteristic of the User/Target dyad. In addition, the Target appears more sensitive to attentional interactions between the User and Tool than either of them do to his interactions with the other.

At least two types of larger-scale patterns have thus far been identified. "Cascades" involve a sequence of triggers, often alternating in source animal, and culminating in some intense level of attentional coordination. In "Suspensions" a prolonged lack of response is maintained by one animal, who remains visually fixated on a particular point (its hand, the grass, etc.), until the partner removes itself from probable engagement (e.g. "closes"), at which point the first finally glances at the partner. Both patterns are most often seen between User and Target. Such interactions provide additional evidence for the functional salience of attentional behavior as well as for the motivational dynamics that characterize social tool use.

Discussion

In summary, we have chosen to conduct a distributed cognition analysis of the triadic primate interaction known as "social tool use". Through this work, we have been able to show that bonobos can and do monitor and respond to the attentional states, including eye gaze, of one another under natural social conditions. Furthermore, rather than relying on a folk-theoretic description of "social tool" that involves attributing humanlike mental states to the animals involved, we have been able to generate an operational definition of social tool. That is, social tool use can be said to occur when one animal (the User) directs some overt social behavior toward a second (the Tool) while keeping its attention primarily fixed on the attentional behavior of a third (the Target) who is likewise attuned to the first.

By examining these interactions in moment-to-moment detail, we can clearly see the co-constructed nature of the cognition involved and the level of complexity that the animals can jointly attain. It would be particularly interesting to do a similar analysis of humans engaged in such interactions, to compare the media, transitional patterns, and levels of sophistication involved. In addition, as a result of this analysis, we can

characterize the individual cognitive abilities of the bonobos not as involving intentionality, deception, or theory of mind, but as reflecting a rudimentary capacity for multi-tasking. Considered one of the most complex of primate behaviors, social tool use is rare even in apes, and thus may be viewed as lying at the limits of those animals' abilities. As such, it suggests that speculation on the critical changes in hominid cognitive evolution might do well to focus on traits that would enable the manipulation of one's own and other's attentional behavior and the capacity to engage in multiple social trajectories simultaneously.

References

Anderson JR, Montant M, Schmitt D (1996) Rhesus monkeys fail to use gaze direction as an experimenter-given cue in an object-choice task. Behavioural Processes 37: 47-56

Ankel-Simons F (1983) A survey of living primates and their anatomy. Macmillan, New York

Bakman R, Gottman JM (1997) Observing interaction: An introduction to sequential analysis (2nd edition). Cambridge University Press, Cambridge.

Baron--Cohen S (1995) Mindblindness: An essay on autism and theory of mind. MIT Press, Cambridge MA London.

Byrne, R, Whiten A (1988) Machiavellian Intelligence, Clarendon Press, Oxford.

Chance MRA (1967) Attention structure as the basis of primate rank orders. Man, 2: 503-518.

Churchland PM (1981) Eliminative materialism and the propositional attitudes. Journal of Philosophy 78:67-90.

Cheney DL, Seyfarth RM, Silk JB (1995) The response of female baboons (*Papio cynocephalus ursinus*) to anomalous social interactions: Evidence for causal reasoning Journal of Comparative Psych 109:134—141.

Coussi--Korbel S, Fragaszy D (1995) On the relation between social dynamics and social learning. Animal Behaviour 50:1441-1453.

de Waal FBM (1986). Dynamics of social relationships. In Smuts, B. B. et al. (eds.), Primate Societies. University of Chicago Press, Chicago.

de Waal FBM, van Hooff J (1981) Side--directed communication and agonistic interactions in chimpanzees. Behaviour 77:164--198.

Fischer KW, Granott N (1995) Beyond one--dimensional change: parallel concurrent socially distributed processes in learning and development. Human Development 38:302—314.

Goodall, J (1986). The Chimpanzees of Gombe, Harvard University Press, Cambridge MA.

Heyes CM (1994) Cues, convergence and a curmudgeon: A reply to Povinelli. Animal Behaviour, 48:242-244.

Humphrey NK (1976) The social function of intellect. In Bateson PPPG, Hinde RA (eds) Growing points in ethology, Cambridge University Press, Cambridge.

Hutchins E (1995) Cognition in the wild. MIT Press, Cambridge MA.

Itakura S, Tanaka M (1998) Use of experimenter-given cues during object-choice tasks by chimpanzees (*Pan troglodytes*), an orangutan *(Pongo pygmaeus*) and human infants (*Homo sapiens*). Journal of Comparative Psychology 112:119-126.

Johnson C.M. (2001) Distributed Primate Cognition: A Review. Animal Cognition 4:167-183.

Jolly A (1985) The evolution of primate behavior, Second edition. Macmillan, New York.

Keating CF, Keating EG (1982) Visual scan patterns of rhesus monkeys viewing faces. Perception 11: 211-219.

Kummer H. (1967) Tripartite relations in Hamadryas baboons. In Altmann, SA (ed) Social Communication Among Primates. Univ. of Chicago Press, Chicago.

Lave J (1988) Cognition in practice: Mind, mathematics and culture in everyday life. Cambridge University Press, Cambridge.

Moore C, Dunham PJ (1995) Joint attention: Its origins and role in development. Erlbaum, Hillsdale NJ.

Peignot P, Anderson JA (1999) Use of experimenter-given manual and facial cues by gorillas (*Gorilla gorilla*) in an object-choice task. Journal of Comparative Psychology 113: 253-260.

Perrett DI, Smith PAJ, Potter, DD, Mistlin AJ, Head AS, Milner AD, Jeeves MA (1985). Visual cells in the temporal cortex sensitive to face and gaze direction. Proc. Roy. Soc. Lond., B, 223:293-317.

Perrett DI, Mistlin AKJ (1990) perception of facial characteristics by monkeys. In Stebbens WC, Berkley MA (eds) Comparative perception. Wiley, NY.

Povinelli DJ (1994) Comparative studies of mental state attribution: A reply to Heyes. Animal Behaviour 48:239-241.

Povinelli DJ, Eddy TJ (1996) Chimpanzees: Joint visual attention. Psychological Science 7: 129-135.

Resnick LB, Levine JM & Teasley SD (1991) Perspectives on socially shared cognition. American Psychological Association, Washington DC.

Rogoff B (1990) Apprenticeship in thinking: Cognitive development in social context. Oxford University Press.

Savage-Rumbaugh ES, Wilkerson BJ (1978) Socio-sexual behavior in *Pan paniscus* and *Pan troglodytes*: A comparative study. J. Human Evol. 7: 327-344.

Silk JB (1999) Male bonnet macaques use information about third party rank relationships to recruit allies. Animal Behaviour 34: 1640—1658.

Smith WJ (1977) The behavior of communicating: An ethological approach. Harvard University Press, Cambridge MA.

Strum S, Forster D, Hutchins E (1997) Why "Machiavellian intelligence" may not be Machiavellian. In: Whiten A, Byrne R (eds) Machiavellian intelligence II. Cambridge University Press, Cambridge New York.

Tomasello M, Call J, Hare B (1998) Five primate species follow the visual gaze of conspecifics. Animal Behaviour 55: 1063-1069.

Vygotsky LS (1978) Mind in society: The development of higher psychological processes. Harvard University Press, Cambridge MA.

Wertsch JV (1985) Culture communication and cognition: Vygotskian perspectives. Cambridge University Press, Cambridge.

Cascade Explains and Informs the Utility of Fading Examples to Problems

Randolph M. Jones (rjones@colby.edu)
Eric S. Fleischman (esfleisc@colby.edu)
Computer Science, Colby College
5847 Mayflower Hill Drive
Waterville, ME 04901-8858 USA

Abstract

Recent research demonstrates that people learn to solve problems more effectively when presented with a series of faded examples and problems, than when presented with completely worked out examples and completely unsolved problems alone. We propose an explanation for this effect, based on the Cascade model. Cascade was originally built to model the self-explanation effect, but it also accounts for other aspects of human learning and problem solving strategies. A relatively straightforward application of Cascade, without alteration, also explains why fading might be beneficial. This explanation provides further support for Cascade as an accurate model of human learning and problem solving, and it also augments the results on fading with specific insight into how, and when, example fading should be effective.

Introduction

There are a variety of research results characterizing factors that facilitate the human ability to learn how to solve problems. Among recent results, Renkl, Atkinson, and Maier (2000) report that people learn more effectively when presented with a set of "faded" examples and study problems than they do when presented with study examples alone (followed by completely unworked problems). Given this result, our interest is in identifying the cognitive mechanisms that explain why the effect exists. Additionally, if we understand the underlying mechanisms, they may provide extra insight into precisely how to create effective sets of study examples and problems. We have identified a potential set of explanatory mechanisms, which are precisely the mechanisms that define Cascade, an existing model of human learning in problem solving (VanLehn, Jones, & Chi, 1991; VanLehn & Jones, 1993a). The Cascade model suggests which study and learning mechanisms contribute to the fading effect, and provides a tool for pinpointing precisely which types of faded examples will be most effective.

Fading Examples

A common curriculum for teaching students to solve problems in a particular domain involves having the students first read some domain material, then study some completely worked out example problems, and finally solve problems that have not been worked out. Various studies explore why it is beneficial for students to study completely worked out examples (e.g., Chi et al., 1989; Pirolli & Anderson, 1985; Renkl, 1997, VanLehn, 1996). The technique of *fading examples* constitutes a particular variation on such a course of study. When fading, the teacher first presents a completely worked example and then follows this with a series of *nearly* completely worked examples. Each subsequent problem removes an explanatory step, forcing the students to solve that portion of the problem themselves. This gives more guidance than a regular problem, because the student still has the guidance of the rest of the worked out example, but it gives less guidance than a completely worked example.

Renkl et al. (2000) demonstrated that different particular methods of presentation can yield different levels of learning. Specifically, they showed that fading sometimes improves learning over other orders of presentation. In their initial study, they presented subjects with a completely worked example, followed by an example that omits the last step in the solution. This was then followed by an example that omits the second to last and last steps and finally one presentation of a completely unworked problem. They compared this situation to one in which students first studied a completely worked example, and then solved a completely unworked problem. The results of this study confirmed that the group of subjects exposed to faded examples learned more effectively than the other group, even though both groups had similar pre-test performances. A more thorough, but similar, experiment produced comparable results.

Renkl et al.'s study provides us with valuable information about potential ways to design study curricula. However, the research also leaves a few unanswered questions. Among these are *why* fading works as well as it does. In addition, it would be useful to know exactly what forms of fading will be effective. Is it always the case that examples should be faded from back to front? Or are there more formal methods we can use to design faded examples? Such questions can, and should, be answered in part by further

experimentation. However, we believe that an existing computer model of human learning in problem solving can also shed light on some of the answers. At worst, the model can help guide future experimentation. At best, the model makes specific predictions and recommendations about how to fade examples.

The Cascade Model

Before presenting Cascade's account of example fading, we will first provide an overview of Cascade's cognitive mechanisms, and how they contribute to explaining other experimental results on human problem solving and learning. Cascade was originally developed to explain the cognitive mechanisms involved in the *self-explanation effect* (Chi et al., 1989; Fergusson-Hessler & de Jong, 1990; Pirolli & Bielaczyc, 1989). Simplifying a bit, the effect shows that people learn more effectively by studying examples when they are careful to explain to themselves as many steps of the example as they can. Students who do not carefully explain worked out example steps do not perform as well on subsequent problems.

Cascade explains this effect as the interaction between two basic learning mechanisms (VanLehn et al., 1991; VanLehn & Jones, 1993b). The basic prediction of Cascade is somewhat intuitive: A student learns effectively when the student is able to identify and patch a specific gap in their knowledge. In Cascade, such knowledge acquisition can occur both while studying worked out examples and while solving problems. When studying an example, Cascade must self-explain each worked step. When the system finds a step it cannot explain, this signifies a knowledge gap. Because the example specifies what the answer is, Cascade can (given appropriate domain-general background knowledge) compensate for the gap by constructing an explanation for the answer. When Cascade successfully creates such an explanation, it learns a new piece of knowledge that it has some faith will work in future problems.

When solving problems, Cascade may also learn by exposing and patching knowledge gaps. However, during problem solving, the process is less constrained because the system does not know the correct answer ahead of time. If the system is missing knowledge, there are potentially a number of ways Cascade could go wrong in attempting to solve a problem. Thus, it is highly likely that the system may expose and attempt to patch a "false" knowledge gap. However, when the system explains examples, it also stores search-control knowledge, essentially remembering the subgoal structure of each example. Experiments with Cascade show that this search-control knowledge can be enough to guide the system into *real* knowledge gaps during problem solving. It can then successfully patch those gaps (VanLehn et al., 1991). Without the search-control knowledge provided by self-explaining examples, Cascade cannot distinguish between real knowledge gaps and otherwise unproductive dead-ends in the attempted problem solution.

Experiments with Cascade have focused mostly on problems involving Newtonian physics, due to the fact that the system's creators had access to physics problem protocol data from Chi et al. The experiments show that Cascade's interacting learning mechanisms account for the basic self-explanation effect, and are able to explain a variety of problem-solving and learning strategies on an aggregate and an individual basis (Jones & VanLehn, 1992; VanLehn et al., 1991; VanLehn & Jones, 1993c).

Because it plays into Cascade's account of example fading, it is worth describing part of the experimental methodology used in the previous Cascade work. The basic method for running experiments with Cascade involves three steps:

1. Configure the system's initial knowledge base, to reflect the hypothetical initial knowledge state of a human subject.
2. Force Cascade to explain exactly those pieces of a series of examples that correspond to observed self-explanations in the subject's protocol (by running it in "explain" mode only on those example pieces).
3. Run Cascade on a series of problems, recording answers, errors, and learning events, to compare with similar events in the encoded subject protocol.

This methodology makes explicit that Cascade does not model all the cognitive process in learning from examples and problem solving. For example, although Cascade can explain the results of self-explanation episodes, it currently has no way to predict which pieces of an example a student will choose to self-explain. However, for the cognitive processes that Cascade models, it does a good job of matching detailed protocol data (Jones & VanLehn, 1992). Thus, we will use the same basic formula to test Cascade's account of example fading.

A Potential Explanation for Fading

Given the background details of how Cascade studies examples and learns to solve problems, we can now sketch the theory behind how Cascade ought to be able to explain the fading effect reported by Renkl et al. It should be clear that, if the Cascade model is accurate, fading of examples can lead to improved learning only if it provides the students with more opportunities to patch their knowledge successfully.

However, it is certainly a valid question whether this is really an accurate characterization of the source of the benefit of example fading. According to Cascade, if a student completely self-explains an example, the student would successfully encounter and patch every potential knowledge gap relevant to that example. Thus, there would be absolutely no benefit to first self-

explaining the example and then combining self-explanation with regular problem solving on a similar subsequent example. The only way example fading could be effective (according to the Cascade model) is if it leads the student to patch a knowledge gap that it would not patch through example studying alone.

As we mentioned above, Cascade can be configured to emulate the self-explanation behavior of any individual subject (Jones & VanLehn, 1992), or to emulate aggregate self-explanation effects (VanLehn et al., 1991). However, it does not explain how or why a subject chooses to self-explain any particular piece of an example. Although an idealized version of Cascade can self-explain a worked example in its entirety, there was not a single subject in Chi et al.'s study who was so thorough in their self-explanation behavior.

Thus, the evidence and model suggest that completely worked examples are never (or at least seldom) as beneficial to students as they could be, because students never (or at least seldom) completely self-explain the examples. This in turn may cause the students to miss opportunities to learn from the examples. However, when a student is given a faded example, there should be some confidence that the student will focus their attention at least on the part of the example that they are requested to solve. Thus, a faded example is similar to a completely worked-out example with a special highlight that says "make sure you self-explain at least this portion of the example".

Our hypothesis is that there are no additional mechanisms required by Cascade to account for the example-fading effect. Rather, using Cascade's existing mechanisms, we hypothesize that example fading forces the student to pay thorough attention to particular portions of the example. Because this forcing is highly focused, the student has a good chance of successfully exposing and patching a knowledge gap (if the faded piece of the example relies on such missing knowledge).

When Cascade is forced to self-explain an individual "faded" portion of an example, we predict that Cascade will acquire new knowledge if that portion of the example exposes a knowledge gap. In such a case, Cascade will exhibit improved learning over normal example studying *if* Cascade was not originally forced to self-explain the same example portion. Due to the fact that we do not have such protocol data from Renkl et al., there is currently no empirical way to tell if this is an accurate characterization of when Renkl et al.'s subjects learned. However, it does provide a testable prediction. Before that experiment is run, however, we should first test the Cascade model to make sure our predictions about its behavior during example fading are accurate.

As an example, a first-year college physics text will often contain an example with a block suspended from a string and sitting on an inclined plane. The example will include some lines that, implicitly or explicitly, involve the inclination of the normal force exerted by the inclined plane on the block. It is not uncommon for a first-year college student never to have heard of normal force, much less to know its inclination. If the student self-explains the entire example, they will come across these portions, reach an impasse, and be forced to create some new understanding concerning normal force. If the student does not bother to self-explain the important parts of the example, they will happily go off and solve inclined plane problems without involving normal force at all (such episodes appear in Chi et al.'s protocol data). Consider taking the same example, but replacing one part of the work with the problem statement, "What is the inclination of the normal force?" This certainly draws the student's attention to normal force, and it explicitly requires the student to understand the concept, if the student is going to be able to finish the problem. In contrast, a student can very easily read a completely worked example without bothering to learn about normal force at all.

Experimental Design

We ran Cascade on Newtonian physics problems, because we already had a thorough task analysis and knowledge representation for the physics domain.

The basic approach involves first using Cascade as a knowledge analysis tool. We ran the system on all the examples and problems given to Chi et al.'s subjects, determining which examples and problems provided the first opportunity to learn a variety of pieces of knowledge. In addition, we used the model to determine which problems required the application of particular pieces of knowledge. This exercise allowed us to create a dependency chart for a number of different knowledge chunks. The chart basically tells us "if you force Cascade to learn this chunk from this example, then it ought to be able to solve this piece of this problem." Or in some more interesting cases, if you force Cascade to learn a particular chunk from an example, it provides the scaffolding for the student to learn *another* chunk in a subsequent problem.

Given this knowledge analysis, it allows us to create experimental runs that simulate effective example fading. The analysis tells us exactly where in each example each knowledge chunk can be learned. Thus, we can focus the system, telling it to explain exactly that portion of the example that will lead to the acquisition of the desired knowledge chunk. This corresponds to creating a focused "faded example", where we allow Cascade not to self-explain most of the example, but force it to self-explain exactly the correct piece. We can then run Cascade on the same example, but with an additional faded portion, so the system can learn a second chunk from the same example. As a control, we can force Cascade to self-explain example

pieces that do not cause any learning, and demonstrate that this has no beneficial effect on future problem solving. An additional control has Cascade "study" the examples without self-explaining the lines, in which case no useful learning occurs.

Results

As predicted, the experiments with Cascade were able to demonstrate improved learning with simulated fading of examples over normal processing of examples. The knowledge analysis showed where Cascade required, or did not require, particular pieces of knowledge to solve a problem. In some cases, the system could solve problems even without any example studying. In other cases, the model was unable to solve problems due to a particular lack of knowledge, even if the system had "studied" an appropriate example without actually self-explaining the key portion of the example. With simulated fading, the system was forced to self-explain the appropriate portions of the examples, and then could solve more problems. Even more, the knowledge analysis was able to tell us exactly which pieces of certain examples should be faded in order to allow the system to solve each particular problem.

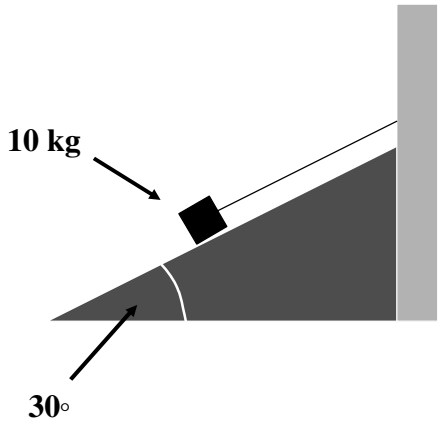


Figure 1: A typical inclined-plane problem. What is the tension in the string?

The experiments we report here concentrated on a thorough analysis of one particular sequence of examples and problems in the Chi et al. study. These examples and problems were all variations of inclined-plane problems (see Figure 1). Given Cascade's initial knowledge base, complete self-explanation of two inclined-plane examples and one "weights and pulleys" example provides the opportunity to expose and patch 12 knowledge gaps.

These opportunities appear in specific portions of the explanation traces for the examples. For example, Cascade can only learn about normal force from self-explaining the normal-force vector in the free-body diagram of the inclined-plane example. If Cascade fails to self-explain this portion of this example, it generates incorrect answers on any future problems that involve normal force. In fact, it usually generates the same incorrect answers as students who also fail to self-explain that portion of that example (Jones & VanLehn, 1992). A portion of the dependency table that covers this chunk appears in Table 1.

Table 1: Dependency chart for the normal-force chunk.

Chunk	`normal-force`
Learning opportunities	Example ixa, line 2 Example ixb, line 2
Used in problems	i1a, i1b, i2a, i2b, i3a, i3b, q5

In one experiment, we forced Cascade to learn about normal force by fading the normal-force portion of the example. Even if Cascade did not self-explain any other part of that example, it learned the normal-force chunk because we forced its attention there.

In each of our experiments, we deleted one of the 12 target rules from Cascade initial knowledge. We then compared Cascade's behavior without any self-explanation to Cascade's behavior when it is forced to self-explain only the portion of the examples that leads it to learn the deleted chunk. The results show that Cascade cannot solve all of the inclined-plane problems without acquiring all 12 of the target pieces of knowledge. It cannot solve *any* of the problems without learning at least some of the 12 chunks (although each of the problems only makes use of a subset of the 12 chunks). More importantly, the results tell us exactly which example portions will lead to success on which problems.

In some cases, Cascade is also able to learn new knowledge chunks during problem solving. In fact, it cannot solve some of the problems without also patching some knowledge gaps during the solution of that problem. The experiments also show that this learning is not always successful unless Cascade has first acquired all the knowledge it needs from the examples. Again, perhaps most informative is that the experiments tell us which pieces of the examples are most beneficial to fade. For example, Cascade cannot solve any problems correctly without first being forced to learn about normal force (as discussed above). But if Cascade is not forced (during example studying) to learn about the inclination of an inclined block's acceleration, it is able to learn that particular piece of knowledge (or at least compensate for it and get the right answer) during problem solving. It can only do this, however, if it has learned about the *existence* of normal force from one of the examples (or from prior experience).

In general, the experimental results confirm our contention that graded fading of examples can map to

improvement in Cascade, if we correspond fading with the forced self-explanation of portions of each example. These results demonstrate the basic dependencies we predicted. Thus, without any alteration to the underlying Cascade model, it is able to provide an explanation for the benefit of faded examples. The key assumption is that example fading serves primarily as a focus of attention, forcing the subject to study closely productive portions of each example. In the closing section, we discuss some implications of this assumption, as well as the possibility of further experiments that would identify more complex interactions between fading and learning.

Conclusions

We feel this work provides two basic contributions. First, it offers additional evidence that Cascade is an accurate model of (at least some of) the cognitive mechanisms involved in studying examples and learning to solve problems. In some ways, this may only be of interest to the designers of Cascade. However, more generally, it allows us to have more faith in using Cascade as a tool both to study human behavior, and perhaps as an aid for curriculum development. In this work, we were able to use Cascade not only as a cognitive model, but also as a knowledge analysis tool. It allowed us to perform a focused dependency analysis that would certainly be beneficial to an instructor creating examples and problems.

However, we should note that much of the hard work that made this analysis possible was performed years ago. We have the benefit of the detailed cognitive analysis of college physics that went into the original design of Cascade. To apply Cascade to any new domains would require a similar intensive effort. However, Cascade at least provides a framework and set of assumptions for creating such knowledge representations. Furthermore, it provides clear principles for where, and how, examples and problems will cause a student to learn target knowledge chunks, as well as which target knowledge chunks contribute to the solution of target test problems.

The second contribution is that we have enhanced our knowledge of how, and why, example fading proves beneficial in some circumstances. The Cascade model suggests that arbitrary fading of examples is not likely to be fruitful, in general. Rather, fading should be focused towards the pieces of examples that will enable learning of a teacher's target knowledge chunks.

Again, this conclusion depends on a key assumption that is integral to Cascade's candidacy as an accurate cognitive model. If Cascade is accurate, it must be the case that faded examples cause effective learning by forcing the student to encounter and overcome an impasse. This is a prediction that can easily be tested. Future experiments on fading should include detailed protocol analysis, and should look for evidence of impasses and learning events in the faded portions of the examples. Such data will confirm or disprove Cascade's account of fading.

It would be prudent to note that the account of fading presented here does not need to be the exclusive source of improved learning. As we mentioned in describing Cascade, the system relies on an interaction between a domain knowledge acquisition learning mechanism and a search-control knowledge learning mechanism. The experiment and explanation reported here relies almost exclusively on the knowledge acquisition mechanism. However, our knowledge of Cascade suggests that it would likely also predict at least some benefit to example fading from the learning of search control knowledge.

Even if a faded portion of an example does not force an impasse and learning event, if it forces some amount of problem solving, Cascade predicts that a student would acquire search control knowledge for the goals and subgoals addressed in the faded portion of the example. As with Cascade's model of the self-explanation effect, such search-control knowledge could benefit later learning, even if no knowledge chunks are learned during the solution of the faded example. However, we have not yet run an experiment along those lines. For now, we will leave that form of learning from a faded example to speculation, to be confirmed or rejected later. It would be most interesting to conduct such computational experiments in concert with similar detailed protocol studies of human subjects. The main point is that Cascade predicts the primary benefit of a faded example is that it forces the student to process parts of the example that they might otherwise ignore. Once a portion of the example is processed, all the cognitive mechanisms that Cascade posits can be brought to bear on learning.

Acknowledgements

This work would not have been possible without the original development of the Cascade system. That, in turn, would not have been possible without the efforts of Kurt VanLehn and Michelene T. H. Chi.

References

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145−182.

Ferguson-Hessler, M. G. M., & de Jong, T. (1990). Studying physics texts: Differences in study processes between good and poor solvers. *Cognition and Instruction, 7*, 41−54.

Jones, R. M., & VanLehn, K. (1992). A fine-grained model of skill acquisition: Fitting Cascade to individual subjects. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 873–878). Hillsdale, NJ: Lawrence Erlbaum.

Pirolli, P., & Anderson, J. R. (1985). The role of learning from examples in the acquisition of recursive programming skills. *Canadian Journal of Psychology, 39*, 240–272.

Pirolli, P., & Bielaczyc, K. (1989). Empirical analyses of self-explanation and transfer in learning to program. In G. M. Olson & E. E. Smith (Eds.), *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 450–457). Hillsdale, NJ: Lawrence Erlbaum.

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science, 21*, 1-29.

Renkl, A., Atkinson, R. K., & Maier, U. H. (2000). From studying examples to solving problems: Fading worked-out solution steps helps learning. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 393–398). Mahwah, NJ: Lawrence Erlbaum.

VanLehn, K. (1996). Cognitive skill acquisition. *Annual Review of Psychology, 47*, 513–539.

VanLehn, K., & Jones, R. M. (1993a). Learning by explaining examples to oneself: A computational model. In S. Chipman & A. L. Meyrowitz (Eds.), *Foundations of knowledge acquisition: Cognitive models of complex learning*. Boston: Kluwer Academic.

VanLehn, K., & Jones, R. M. (1993b). Integration of explanation-based learning of correctness and analogical search control. In S. Minton (Ed.), *Machine learning methods for planning*. Los Altos, CA: Morgan Kaufmann.

VanLehn, K., & Jones, R. M. (1993c). Better learners use analogical problem solving sparingly. *Machine Learning: Proceedings of the Tenth International Conference* (pp. 338–345). San Mateo, CA: Morgan Kaufmann.

VanLehn, K., Jones, R. M., & Chi, M. T. H. (1991). A model of the self-explanation effect. *Journal of the Learning Sciences, 2*, 1–59.

Modelling the Detailed Pattern of SRT Sequence Learning

F.W. Jones (psm1fj@surrey.ac.uk)[1]
University of Cambridge, Department of Experimental Psychology,
Downing Street, Cambridge, CB2 3EB, UK.

I.P.L. McLaren (iplm2@cus.cam.ac.uk)
University of Cambridge, Department of Experimental Psychology,
Downing Street, Cambridge, CB2 3EB, UK.

Abstract

Any viable model of serial reaction time (SRT) sequence learning needs to be able to capture the relative extents to which participants learn different SRT sub-sequences. However, previous attempts to empirically establish a relative pattern of this sort have failed to fully control for 'sequential-effects', or have not satisfactorily ruled-out the influence of speed-accuracy trade-off. In this paper it is shown that, when sequential-effects are controlled for in a two-choice SRT task, participants learn those sub-sequences that end in an alternation better than those that end in a repetition. It is further demonstrated that, at least for the parameters investigated, a buffer network, a Jordan network, an SRN, an augmented SRN, and an AARN are unable to account for this pattern.

Introduction

Since its introduction (Lewicki, Czyzewska & Hoffman, 1987; Nissen & Bullemer, 1987) the serial reaction time (SRT) task has proved a popular way to assay human sequence learning. In the most common version of this task a stimulus can appear in one of several locations on a computer screen, each of which has a corresponding response key. Whenever a stimulus appears, participants simply have to press the appropriate key as quickly and accurately as possible. Crucially, usually unknown to the participants, the order of locations in which the stimulus is presented follows a sequence, at least for the majority of the time.

Typically, participants' reaction times (RTs) on trials that are consistent with this sequence become significantly faster, either than their RTs on occasional inconsistent probe trials, or than the RTs of a control group that is only exposed to a (pseudo-)random ordering (e.g. Anastasopoulou & Harvey, 1999). This is taken to imply that the participants have learnt at least part of the sequence, and that they are using this information to prepare for the stimulus or response on the next trial.

The widespread use of this task makes it a priority to develop an adequate model of human SRT performance. A particularly stringent test of any candidate model is to capture the relative extents to which participants learn different parts (sub-sequences) of an SRT sequence. Unfortunately, however, nearly all of the patterns of sub-sequence learning that have been reported in the literature are potentially confounded by 'sequential-effects' (cf. Anastaspoulou & Harvey, 1999; Shanks & Johnstone, 1999).

'Sequential-effects' can be defined as the influence that the previous series of stimulus/response locations has on the participant's current response, in a choice-RT task in which the trial order is (pseudo-)random (e.g. Soetens, Boer & Hueting, 1985). To illustrate, in a two-choice task at a response-stimulus-interval (RSI) of 50ms, participants tend to respond faster when the stimulus appears in the same location as it did on the prior trial (Soetens et al., 1985).

Therefore, in order to more accurately assess SRT sequence learning, the performance of the sequence group and pseudo-random control group need to be compared on a 'test-phase' in which they are both exposed to the same order of trials; the assumption being that the two groups should manifest the same sequential-effects when responding to identical trial orders, allowing the difference between the groups to be used as an index of sequence learning.

However, the majority of previous SRT studies have failed to control for sequential effects in this manner (e.g. Nissen & Bullemer, 1987), calling into question the apparent patterns of sub-sequence learning they report.[2] Furthermore, the two previous studies that have adequately controlled for sequential effects, namely Anastasopoulou and Harvey (1999) and Shanks and Johnstone (1999), have not reported error data in sufficient detail to rule out the possibility that participants traded speed and accuracy between different sub-sequences.

Therefore it was necessary to carry out a new experiment in order to establish a relative pattern of SRT sub-sequence learning against which models of the SRT task could be tested.

Experiment

The experiment comprised a five session two-choice SRT task, with an Experimental Group that was trained on four sub-sequences and a Control Group that was exposed to a pseudo-random ordering for the same number of trials. To control for sequential effects, following training all

[1] Now at: University of Surrey, Department of Psychology (Clinical Psychology PsychD), Guildford, Surrey GU2 7XH, UK.

[2] Some authors have attempted to reduce the influence of sequential-effects by removing trials that they feel are particularly contaminated (e.g. Jimenez, Mendez & Cleeremans, 1996). However, it is not clear that this is sufficient.

participants were 'tested' on pseudo-random orderings, with the difference between the two groups being taken as an index of sequence learning.

Method

There is not the space to give full methodological details, but these will be provided in Jones and McLaren (in prep.).[3]

Participants, Stimuli and Apparatus The participants were 24 Cambridge University students and members of the public, whose ages ranged between 18 and 47. 16 were randomly assigned to the Experimental Group and 8 to the Control. This difference in numbers did not cause any reliable differences in variance between the groups; therefore the subsequent ANOVAs are valid. The experiment was run on a Macintosh LCIII, and the stimulus comprised a circle 1.9 cm in diameter which could appear 2.2 cm to the right or left of the centre of the screen. The two response keys were spatially compatible with these locations.

Trial Order The trials were batched into blocks of 120, and are described here in terms of Xs and Ys. The assignment of X and Y to right and left was counterbalanced across participants. As illustrated in Figure 1, the building blocks of the 'sequence-blocks' were four sub-sequences, each three trials long; specifically XXX, XYY, YYX and YXY. These sub-sequences conform to the rule 'if the first and second trials are the same then the third is an X, but if they differ it is a Y'. Sequence-blocks were constructed by concatenating 10 of each of these sub-sequences in a random order. Thus every third trial in a sequence-block was predictable on the basis of the previous two. Participants were not informed about the special status of third trials, and the RSI following third trials was the same as that following other trials.[4] It is also worth noting that when the properties of these sequence-blocks are considered on a trial-by-trial basis, then approximately two-thirds of the trials are consistent with the rule, and on average the four sub-sequences occur equally frequently. (This was confirmed by the Monte-Carlo method.)

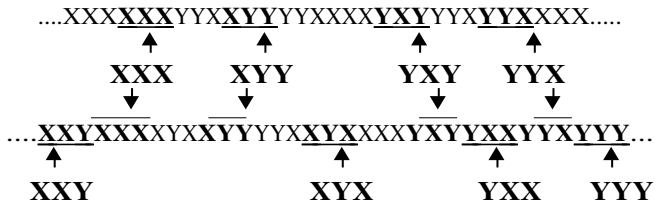


Figure 1: Examples of portions of a sequence-block (top) and a pseudo-random-block (bottom), with some of the constituent triplets highlighted.

[3] This paper will not include the simulation work presented here.

[4] The sequence-blocks were constructed so that the third trials had a special status because the original purpose of the experiment was to compare the pattern of sub-sequence learning between this incidental condition and one in which participants were told about the third trials (see Jones & McLaren, in prep.).

'Pseudo-random-blocks' comprised 5 of each of the sub-sequences (XXX, XYY, YYX and YXY) and 5 of each of the complementary triplets with alternate endings (XXY, XYX, YYY and YXX), randomly concatenated.

Design Each session comprised 20 blocks. The first 10 blocks of session one and the last 10 blocks of session five constituted the pre- and post-training 'test-phases', and were formed of pseudo-random-blocks for both groups. The 80 block 'training-phase' in between these comprised sequence-blocks for the Experimental Group and pseudo-random-blocks for the Control Group.

Procedure The procedure followed that described for the standard SRT task in the introduction, with the addition that participants were paid a performance bonus, designed to encourage them to be as fast and accurate as possible. The RSI was 500ms.

Results and Discussion

Due to space constraints, only those analyses most relevant to the subsequent modelling are reported here, but see Jones and McLaren (in prep.) for full details. Furthermore, only the RT results have been presented because the same trends were observed in the errors, ruling out speed-accuracy trade-off.

To assay sub-sequence learning, the data from the third trials of the first five blocks[5] of the post-training test-phase were divided up with respect to the identity of the previous two trials (i.e. XX, XY, YY and YX). They were then further divided on the basis of whether the third trial was consistent with the sub-sequence begun by the previous two trials (e.g. XXX) or inconsistent with it (e.g. XXY). For the Control Group consistency was a dummy variable, since they had never been trained on the sub-sequences.

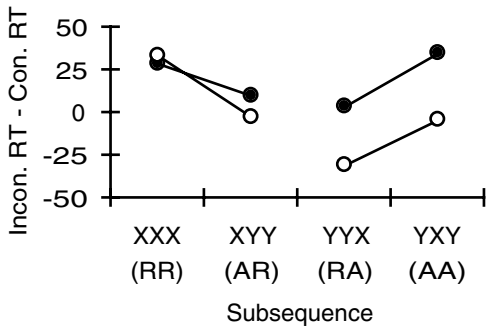


Figure 2: The mean difference scores from the third trials of the first half of the post-training test phase. Filled circles = Experimental Group. Open circles = Control Group.

A mean RT for each of the resulting eight trial types was calculated, on a per participant basis. To reduce variability, RTs were not taken from error trials or trials following an error. Then, for each sub-sequence, a difference score was constructed by subtracting the consistent RT from the inconsistent one. This measure was used because it subtracts

[5] An analysis of the learning curves suggested that extinction had set in by blocks 6-10 (see Jones & McLaren, in prep.).

out individual variability in baseline RT. The mean difference score for each sub-sequence, per group, are shown in Figure 2.

Following the convention adopted in the two-choice sequential effects literature (e.g. Soetens et al., 1985), the four sub-sequences were coded in terms of whether each trial was a repetition (R) or an alternation (A) of the previous trial (i.e. XXX=RR, XYY=AR, YYX=RA and YXY=AA). The difference scores were then analysed using an ANOVA with the factors: group (experimental vs. control), first position in the alternation-repetition sub-sequence code (A or R), and second position (A or R).

This demonstrated that the Experimental Group's difference scores were significantly larger those of the Control Group ($F(1,22)=31.77$, $p<.01$). This suggests that the participants in the Experimental Group learnt at least some of the sequential contingencies, because sequence learning should produce slower RTs on inconsistent trials and faster RTs on consistent trials, and thus a larger inconsistent minus consistent score.

The group X first-position (reading from left to right) interaction was not significant ($F<1$), nor was the group X first-position X second-position interaction ($F<1$). However, the group X second-position interaction was reliable ($F(1,22)=5.14$, $p<.05$). According to a simple effects analysis, this interaction arose because the Experimental Group's difference scores were significantly greater than Control for those sub-sequences that ended in an alternation ($F(1,22)=18.12$, $p<.01$) but not for those that ended in a repetition ($F(1,22)=0.17$, $p>.1$).

Thus it would appear that differential learning of the sub-sequences occurred, with participants only learning those sub-sequences that ended in an alternation within the time of the experiment.

Furthermore, this pattern would appear not to be contaminated by a floor effect, because if just the results from the inconsistent trials are analysed then the same trend is observed. (The expression of learning cannot be masked by RT being at floor on inconsistent trials because learning should slow responding on such trials.)

Finally, when data from all trials was included in the analysis, rather than just from the third trials, a similar pattern of sub-sequence learning was observed. Thus there was no evidence to suggest that the participants in the Experimental Group had learned about the special status of third trials (i.e. that they were always consistent during training). Rather, participants appear to have learnt the 2/3 contingencies that are present on a trial by trial basis (see method section).

In summary, in a two choice SRT task in which sequential effects are controlled for, it appears that people learn those sub-sequences that end in an alternation better than those that end in a repetition. But can the current neural network models of SRT sequence learning capture this pattern?

Modelling

To address this question, a variety of different neural network models of SRT sequence learning were presented with an exact analogue of the human task. These models all include some form of memory for previous trials. It is by associating a combination of the contents of this memory and a representation of the current trial with the identity of the next trial that they learn the sequential contingencies. As the models learn they become more accurate at predicting the next trial's identity, and consequently the mean squared error (MSE) of their prediction reduces.

Therefore, in the following simulations the MSE has been taken as an index of the models' RT to the stimulus that it is attempting to predict; the rationale being that a low MSE indicates that a model accurately expects the location of the next stimulus and so it should react more quickly to it when it occurs. While some authors transform the MSE using a decision rule or mechanism (cf. Cleeremans, 1993), this was avoided in order to prevent the properties of the models from potentially being obscured by the addition of an extra process.

In all the following simulations a localist code, with one unit for X and one for Y, was used to represent the input to the models and their predictions. The trial order was generated in exactly the same way as in the experiment, and for each type of model both an Experimental and Control Group were run. Finally, to assay sequence learning the models' MSEs from the first half of the post-training test-phase were analysed in the same way as the human data.

The different types of networks studied will now be considered in turn.

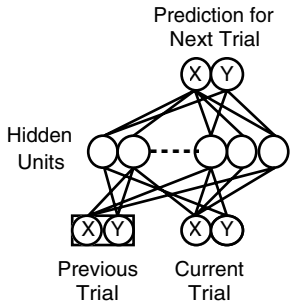


Figure 3: A buffer network.

The Buffer Network

The architecture of a buffer network is shown in Figure 3. In this model a memory is simply implemented by presenting as input to the network all the necessary elements of the sequence. Thus, the network's input comprises a representation of the current trial and a decayed (by half) representation of the previous trial. Trials prior to this are not included since the contingencies in the sequence-blocks are only depended upon the previous two trials. The architecture includes a layer of hidden units, because these enable it to learn non-linearly separable mappings (Rumelhart, Hinton & Williams, 1986), and the weights are updated using the backpropagation algorithm (Rumelhart, Hinton & Williams, 1986). For more details concerning buffer networks see Cleeremans (1993, pp. 141-143). (Note, a momentum term was not employed).

Thirty-two separate buffer networks, each with 4 hidden units, were run on an analogue of the experiment. Sixteen formed the Control Group and 16 the Experimental Group.

Each network 'subject' had a different set of randomly initialised weights. And, the learning rate was set to 0.8, since this value meant that the networks had learnt two of the four sub-sequences by the post-training test-phase, like the human participants had.

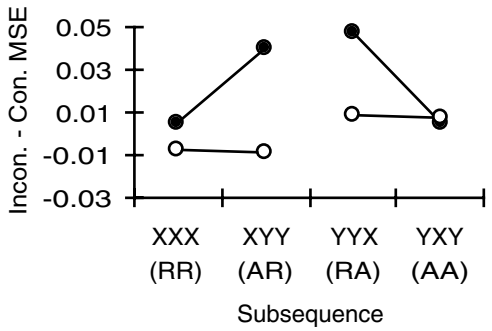


Figure 3: The Buffer networks' difference scores from the third trials of the first half of the post-training test phase. Filled circles = Experimental Group. Open circles = Control.

The mean post-training MSE difference scores are shown in Figure 3. The data were analysed using a group (2 levels) by sub-sequence (4 levels) ANOVA. This revealed that the Experimental Group's scores were significantly higher than Control ($F(1,30)=17.23$, $p<.01$), indicating that they had acquired some of the sequential contingencies. Moreover, there was evidence of differential sub-sequence learning (group X sub-sequence: epsilon corrected $F(2,57)=5.41$, $p<.05$); with a simple-effects analysis demonstrating that subjects had reliably learnt subsequences XYY/AR and YYX/RA (respectively $F(1,30)=22.05$, $p<.01$; $F(1,30)=5.84$, $p<.05$), marginally learnt XXX/RR ($F(1,30)=4.06$, $.1>p>.05$), and not learnt YYX/AA ($F(1,30)=0.24$, $p>.1$). The model made similar predictions with 20 hidden units.

However, these predictions differ from the pattern expressed by human participants (i.e. stronger learning of those sub-sequences that ended in an alternation). Therefore, the buffer network was unable to model the human data, at least with 4 or 20 hidden units.

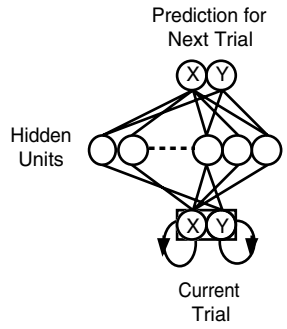


Figure 4: A Jordan network.

The Jordan Network

The architecture of a Jordan network (Jordan, 1986), modified to model the SRT task (Cleeremans, 1993, pp. 139-141), is illustrated in Figure 4. In this model the identity of previous trials are not explicitly represented on separate pools of input units. Rather, a memory of the sequence is implemented by adding to each input unit a self-recurrent connection of fixed weight, which is 0.5 in this case. As with the buffer network the variable weights are updated using the backpropagation algorithm.

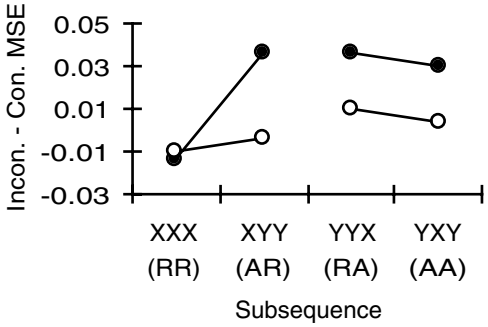


Figure 5: The Jordan networks' difference scores from the third trials of the first half of the post-training test phase.

To provide reliable results, 34 network subjects were run in both groups. Each network had 4 hidden units and a learning rate of 0.5. The results are shown in Figure 5. An ANOVA revealed reliable evidence of sequence learning ($F(1,66)=42.92$, $p<.01$) and of differential sub-sequence learning ($F(3,198)=17.28$, $p<.01$). A follow up simple-effects analysis demonstrated that the networks had learnt all the sub-sequences except XXX/RR (XXX/RR: $F(1,66)=1.47$, $p>.1$; XYY/AR: $F(1,66)=39.60$, $p<.01$; YYX/RA: $F(1,66)=20.23$, $p<.01$; YXY/AA: $F(1,66)=33.38$, $p<.01$). Therefore, with 4 hidden units, the Jordan network achieved a closer match to the human data than the buffer network, but it still was unable to capture the full detail of the pattern. When the number of hidden units was increased to 20, the networks' results were less similar to the human data.

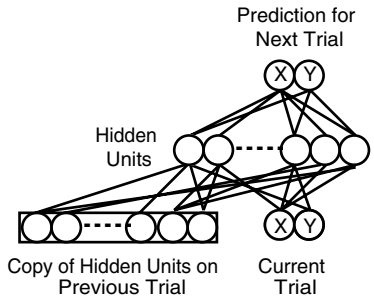


Figure 6: An SRN.

The Simple Recurrent Network (SRN)

Figure 6 illustrates the architecture of the SRN, which was developed by Elman (1990). In the SRN a memory of previous items in the sequence is implemented by providing

as additional input to the network its own hidden unit activations from the previous trial. As with the other models, the weights are updated by the backpropagation algorithm.

With 4 or 20 hidden units, and a range of large learning rates (0.5, 0.8, and 1.0), the SRN was incapable of reliably learning any of the sequential contingencies by the post-training test-phase ($p>.1$). However, when sixteen 40 hidden unit SRNs were run in each group (learning rate 0.5), an ANOVA revealed that reliable sequence learning did occur ($F(1,30)=13.41$, $p<.01$). For the means see Figure 7.

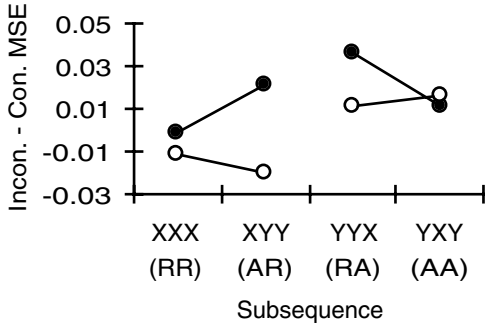


Figure 7: The SRNs' difference scores from the third trials of the first half of the post-training test phase.

Moreover, there was reliable evidence of differential sub-sequence learning ($F(3,90)=8.91$, $p<.01$), and according to a simple-effects analysis the network subjects had learnt all the sub-sequences except YXY/AA (XXX/RR: $F(1,30)=4.97$, $p<.05$; XYY/AR: $F(1,30)=13.82$, $p<.01$; YYX/RA: $F(1,30)=9.37$, $p<.01$; YXY/AA: $F(1,30)=1.59$, $p>.1$). However, this pattern deviates from the advantage for those sub-sequences that ended in an alternation seen in the experiment.

The Augmented SRN

In order to allow the SRN to capture a higher proportion of the variance in their SRT data, Cleeremans and McClelland (1991) made two modification to it, producing the Augmented SRN. First, to model the short term priming effect of previous learning episodes, weights and biases were divided into both a fast and slow component.

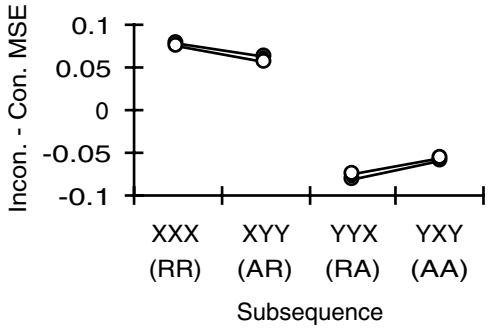


Figure 8: The Augmented SRNs' difference scores from the third trials of the first half of the post-training test phase.

Second, to capture response repetition effects the model's prediction of the next response was made dependent upon a decaying trace of previous responses as well as the SRN's output. For more details see Cleeremans and McClelland (1991).

To determine whether this model could capture the experimental results, 16 networks with 4 hidden units were run in each group. The learning rates were 1.3 and 1.0 for the fast and slow weights respectively. The other parameters were as described in Cleeremans and McClelland (1991). Given that the SRN's output activations are transformed in this model, the transformed activations were employed in the calculation of the MSE.

The results are shown Figure 8. While the trends may appear to be very small, an ANOVA did reveal that the networks had reliably learnt at least some of the contingencies ($F(1,30)=15.64$, $p<.01$) and that this learning varied across the sub-sequences ($F(3,90)=38.64$, $p<.01$). A subsequent simple-effects analysis demonstrated that XXX/RR and XYY/AR were learnt ($F(1,30)=28.17$ and 58.01, $p<.01$), while the Experimental Group's scores were significantly lower than Control for YYX/RA and YXY/AA ($F(1,30)=51.41$ and 11.63, $p<.01$). Thus the pattern of subsequence learning was the nearly opposite to that expressed by human subjects. Nor did increasing the number of hidden units to 20 substantially improve the situation.

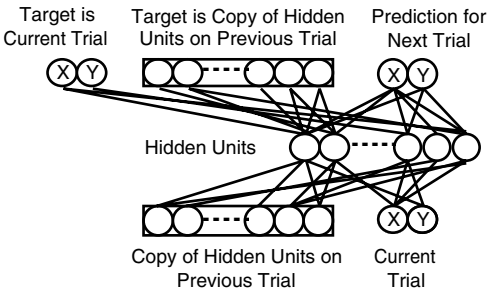


Figure 9: An AARN.

The Autoassociative Recurrent Network (AARN)

The AARN, which was developed by Maskara and Noetzel (1993), is shown in Figure 9. It differs from a standard SRN in that the network is taught not only to predict the next trial, but also to predict the pattern of activity across the input layer. This is implemented by including a extra pool of output units, with each of these units corresponding to a particular unit in the input layer. The target for one of these new output units is the activity of its corresponding input unit on the same trial.

Fifty AARNs, with four hidden units and a learning rate of 0.8, were run in each group. This large number of subjects was required to produce significant results. It is also worth noting that only the activations of the two output units that comprised the network's prediction for the next trial were included in the calculation of the MSE.

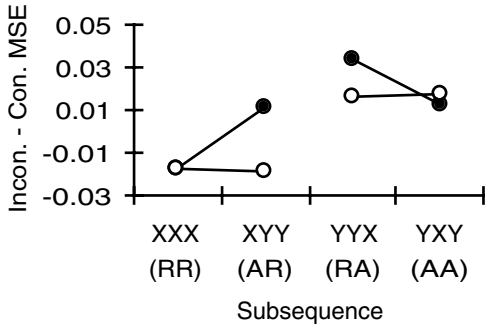


Figure 10: The AARNs' difference scores from the third trials of the first half of the post-training test phase.

The networks' mean MSE difference scores are shown in Figure 10. An ANOVA revealed that the Experimental Group had learnt at least some of the sequential contingencies ($F(1,98)=47.29$, $p<.01$) and that the sub-sequences had been learnt differentially ($F(3,294)=9.77$, $p<.01$). According to a simple effects analysis, subsequences XYY/AR and YYX/RA were learnt ($F(1,98)=39.81$ and 11.59, $p<.01$), while the remaining two were not ($F(1,98)=0.04$ and 1.37, $p>.1$). This pattern contrasted with the stronger learning of those sub-sequences that ended in an alternation observed in the human data. Furthermore, changing the number of hidden units to 20 did not improve the situation.

Dominey's Network

The ability of Dominey's (1995) model to capture the human data has also been examined. However, thus far we have been unable to find a set of parameters that enables the model to display any evidence of sequence learning on the task.

General Discussion

To summarise, previous attempts to establish a relative pattern of SRT sub-sequence learning are flawed because either they fail to control for sequential effects or do not present error data in sufficient detail. In a two-choice SRT task designed to overcome these flaws, it was found that people learn those sub-sequences that end in an alternation better than those that end in a repetition. Surprisingly, however, a buffer network, the Jordan network, the SRN, the augmented SRN and the AARN all appear unable to capture this result, at least with the parameters investigated.

A critic could argue that our human data might be beyond the scope of these models because it may reflect 'explicit' rather than 'implicit' learning. However, other work in our laboratory suggests that people show a different pattern when learning explicitly, namely they find XXX/RR the most salient sub-sequence. Therefore, it would seem that current neural network models of SRT sequence learning will probably at least need to be modified in order to accommodate the data presented here.

Acknowledgements

This research was funded by an MRC research studentship awarded to F.W. Jones. The authors would thank Stephen Monsell, David Shanks, Nick Yeung and two anonymous reviewers for their helpful comments.

References

Anastasopoulou, T., & Harvey, N. (1999). Assessing sequential knowledge through performance measures: the influence of short-term sequential effects. *Quarterly Journal of Experimental Psychology, 52A*, 423-448.

Cleeremans, A. (1993). *Mechanisms of implicit learning*. Cambridge, MA: MIT Press.

Cleeremans, A. & McClelland, J.L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General, 120*, 235-253.

Dominey, P.F. (1995). Complex sensory-motor sequence learning based on recurrent state representation and reinforcement learning. *Biological Cybernetics, 73*, 265-274.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science, 14*, 179-211.

Jimenez, L., Mendez, C., & Cleeremans, A. (1996). Comparing direct and indirect measures of sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 948-969.

Jordan, M.I. (1986). Attractor dynamics and parallelism in a connectionist sequential machine. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ. LEA.

Lewicki, P., Czyzewska, M., & Hoffman, H. (1987). Unconscious acquisition of complex procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 523-530.

Maskara, A., & Noetzel, A. (1993). Sequence recognition with recurrent neural networks. *Connection Science, 5*, 139-152.

Nissen, M.J., & Bullemer, P. (1987). Attentional requirements of learning: evidence from performance measures. *Cognitive Psychology, 19*, 1-32.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart and J.L. McClelland (Eds.), *Parallel Distributed Processing. Volume 1*. Cambridge, MA: Bradford Books.

Shanks, D.R., & Johnstone, T. (1999). Evaluating the relationship between explicit and implicit knowledge in a sequential reaction time task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1435-1451.

Soetens, E., Boer, L.C., & Hueting, J.E. (1985). Expectancy or automatic facilitation? Separating sequential effects in two-choice reaction time. *Journal of Experimental Psychology: Human Perception and Performance, 11*, 598-616.

Where Do Probability Judgments Come From? Evidence for Similarity–Graded Probability

Peter Juslin (peter.juslin@psy.umu.se)
Department of Psychology, Umeå University
SE-901 87, Umeå, Sweden

Håkan Nilsson (hakan.nilsson@97.polmag.umu.se)
Department of Psychology, Umeå University
SE-901 87, Umeå, Sweden

Henrik Olsson (henrik.olsson@psy.umu.se)
Department of Psychology, Umeå University
SE-901 87, Umeå, Sweden

Abstract

This paper compares four models of the processes and representations in probability judgment. The models represent three principles that have been proposed in the literature: 1) the *representativeness heuristic* (interpreted as relative likelihood or prototype-similarity), 2) *cue-based relative frequency*, and 3) *similarity-graded probability*. An experiment examined if these models account for the probability judgments in a category learning task. The results indicated superior overall fit for similarity-graded probability throughout training. In the final block, all models except similarity-graded probability were refuted by data.

Introduction

Where do probability judgments come from? This question has been fiercely debated the last decades in research on judgment under uncertainty. In the late sixties the conclusion was that probability judgments are fairly accurate reflections of *extensional* properties of the environment such as frequencies (Peterson & Beach, 1967). This changed with the influential *heuristics and biases* program in the seventies and eighties, which emphasized that probability judgments are guided by *intensional* aspects like similarity (Kahneman, Slovic, & Tversky, 1982). The nineties saw a renewed interest in the idea that extensional properties are reflected in peoples' probability judgments as specified by the *ecological models* (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, 1994). A third alternative combines intensional and extensional properties in an exemplar model to produce *similarity-graded probabilities* (Juslin & Persson, 2000).

Only rarely have these accounts been contrasted in studies that chart the processes and representations that underlie probability judgments. We compare four models of how people make probability judgments in a category learning task. The task involves assessment of the probability that a probe with feature pattern t belongs to one of two mutually exclusive categories, A or B. For example, a physician may assess the probability that a patient with symptom pattern t suffers from one of two diseases. The models represent three principles that have been proposed in the judgment literature: the *representativeness heuristic* (two versions), *cue-based relative frequency*, and *similarity-graded probability*. We present a category structure that allows us to contrast predictions derived from these hypotheses.

Representativeness Heuristic

According to the representativeness heuristic, people judge the probability that an object or event belongs to a category on the basis of the degree to which it is representative of the category, or reflects salient features of the process that generated it (Kahneman et al., 1982). The representativeness heuristic is routinely evoked post hoc to explain cognitive biases but has not been subjected to careful tests in inductive learning tasks.

A *relative-likelihood* interpretation of representativeness states that the probability judgment $p(A)$ that probe t belongs to A is made by comparing the likelihood of t in category A relative to its likelihood in categories A and B:

$$p(A) = \frac{.5d + f(t|A)}{d + f(t|A) + f(t|B)}, \qquad \text{Eq. 1}$$

where $f(t|A)$ and $f(t|B)$ are the *relative frequencies* of feature patterns identical to t in categories A and B, respectively. To allow for pre-asymptotic learning (Nosofsky, Kruschke, & McKinley, 1992) and response error in the use of the overt probability scale (Erev, Wallsten, & Budescu, 1994), all models in this paper are equipped with a free parameter d for dampening. The dampening effectively pulls the predictions towards .5 (e.g., an un-dampened prediction of 1 becomes somewhat less extreme as a result of d). Eq. 1 implies that the probability judgment that, say, a patient with symptom pattern t has disease A is a direct function of

the likelihood of these symptoms given disease *A*.[1]

A *prototype* interpretation of representativeness is that the probability judgments derive from the *similarities* $S(t|P_A)$ and $S(t|P_B)$ of *t* to the category prototypes P_A and P_B, respectively:

$$p(A) = \frac{.5d + S(t|P_A)}{d + S(t|P_A) + S(t|P_B)}, \qquad \text{Eq. 2}$$

where the similarity is computed by the multiplicative similarity rule of the context model (Medin & Schaffer, 1978),

$$S(t, y) = \prod_{j=1}^{D} d_j \ , \ d_j = \begin{cases} 1 \ if \ t_{.j} = y_{.j} \\ s \ if \ t_{.j} \neq y_{.j} \end{cases}, \qquad \text{Eq. 3}$$

where *y* is a prototype (as in Eq. 2 above) or an exemplar (as in Eq. 5 below). The value of d_j is 1 if the values on feature *j* match and *s* if they mismatch. *Similarity s* is a free parameter in the interval [0, 1] for the impact of mismatching features.

On this view, the probability judgment that a patient with symptom pattern *t* has disease *A* is a function of *t*'s similarity to the prototypical symptom pattern for disease *A*. The prototype is defined by the modal (i.e., most frequent) feature value in the category on each feature dimension. When the feature values are equally common, we selected the feature value that generated the more frequent overall pattern in the category.

Cue-Based Relative Frequency

The idea that probability judgments derive from cue-based relative frequency is represented by *Probabilistic Mental Model theory (PMM-theory;* Gigerenzer et al., 1991; see e.g., Juslin, 1994, for similar ideas). These ideas have been used to scaffold global predictions in studies of realism of confidence, but not been tested in studies of inductive learning.

In the current context, we interpret PMM-theory as suggesting that the probability judgment that probe *t* belongs to category *A* is a function of the cue value (α_1) of the single most valid cue that can be applied:

$$p(A) = \frac{.5d + F(A|\alpha_1 ***)}{d + F(A|\alpha_1 ***) + F(B|\alpha_1 ***)}, \qquad \text{Eq. 4}$$

where $F(A|\alpha_1***)$ and $F(B|\alpha_1***)$ are the *frequencies* of category *A* and *B* exemplars with cue value α_1, respectively, and the symbol "∗" denotes that the other cue values are discarded (there are four features in the experiment presented below). Eq. 3 represents the relative frequency of category *A* conditional on presence of cue value α_1. Thus, a subjective probability judgment is a reflection of the validity of the cue with the highest cue-validity that is present in the event or object being judged. This strategy is known as *Take The Best* (TTB) meaning that you rely on the cue with the highest validity (Gigerenzer, Todd, & the ABC Group, 1999).

Similarity-Graded Probability

A class of models that combines intensional and extensional aspects is exemplar models in categorization research. In exemplar models, decisions are made by comparing new objects with exemplars stored in memory. The *context model* (Medin & Schaffer, 1978) responds to both similarity (intensional property) and frequency (extensional property) in general, and to only one of these factors in predictable circumstances (Juslin & Persson, 2000). PROBEX (i.e., PROBabilities from EXemplars; Juslin & Persson, 2000) is a model of probability judgment based on the context model.

With PROBEX, probability judgments are made by comparisons between the probe *t* and retrieved exemplars x_i (i = 1…I). The exemplars are represented as vectors of *D* features (in the present experiment, *D*=4 and the features are binary). Continuing with the example of medical diagnosis, a patient with symptom pattern *t* leads to retrieval of stored exemplars of previous patients with similar symptoms and their diagnoses. The probability judgment is a weighted average of the outcome indices $c(x_i)$ for the exemplars, where $c(x_i)$=1 for exemplars in category *A* and $c(x_i)$=0 for exemplars in category *B*. The weights in the average are the respective probe-exemplar similarities $S(t|x_i)$:

$$p(A) = \frac{.5d + \sum_I S(t|x_i)c(x_i)}{d + \sum_I S(t|x_i)}, \qquad \text{Eq. 5}$$

where similarity is computed from Eq. 3. This hypothesis implies that if a new patient with symptom pattern *t* is similar to many exemplars x_i with diagnosis *A*, the probability that the new patient has disease *A* is high.

The complete version of PROBEX involves a sequential sampling of exemplars, but this aspect is ignored in the present application. This effectively reduces Eq. 5 to the original context model (Medin & Schaffer, 1978) with a dampening (see Nosofsky et al., 1992, for a similar formulation), but with one crucial difference: $p(A)$ does not refer to a predicted proportion of category *A* classifications, but to a prediction of a probability judgment.

With similarity parameter *s*=0, only exemplars with feature patterns identical to *t* affect the judgment and Eq. 5 emulates a "*picky frequentist*" (Juslin & Persson, 2000).[2] Ignoring the dampening *d*, Eq. 5 then computes

[1] "Direct function" means that the predicted probability judgments are a function of likelihoods alone, not likelihoods *and* prior probabilities, as implied by Bayes' theorem.

[2] This version of Eq. 5 is formally identical to Bayesian estimation of a probability with the Beta-distribution and parameters α and β equal to $.5d$.

the relative frequency of category *A* among exemplars with identical features. For $s>0$, Eq. 5 computes a *similarity-graded probability* that is both affected by the frequency of exemplars, and the probe-exemplar similarities. Note that, although PROBEX responds to similarity, it is not identical to the representativeness heuristic. For example, PROBEX (Eq. 5) cannot produce a conjunction fallacy, unless amended with auxiliary assumptions of some sort (Juslin & Persson, 2000). PROBEX has been fitted to people's probability judgments in a general knowledge task (Juslin & Persson, 2000) but not been tested in inductive learning tasks.

Category Structure and Predictions

The problem with contrasting these three hypotheses is that in most category structures, they generate highly correlated predictions. Table 1, however, provides one category structure that implies qualitatively distinct predictions for certain critical exemplars (Figure 1).

Table 1: The categories with the 20 x 3 exemplars.

X	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	Category		
1	1	1	1	1	3	3	3	3	5	5	5	5	A	A	A
2	1	1	1	1	3	3	3	3	5	5	5	5	A	A	A
3	1	1	1	1	3	3	3	3	5	5	5	5	A	A	A
4	1	1	1	1	3	3	3	3	5	5	5	5	A	A	B
5	1	1	1	1	3	3	3	3	5	5	5	5	A	B	B
6	1	0	0	0	3	2	2	2	5	5	5	4	A	A	A
7	1	0	0	0	3	2	2	2	4	4	4	4	A	A	B
8	1	0	0	0	3	2	2	2	4	5	4	4	A	A	B
9	0	0	0	0	3	2	2	2	4	4	4	4	A	A	B
10	0	0	0	0	3	2	2	2	4	5	4	4	A	A	B
11	1	1	0	0	3	2	2	2	4	4	4	4	B	A	B
12	1	1	0	0	3	2	2	2	4	4	5	4	B	A	B
13	0	1	0	0	3	2	2	2	4	4	4	4	B	A	B
14	0	1	0	0	3	2	2	2	4	4	5	4	B	A	B
15	0	1	0	0	3	2	2	2	4	4	4	4	B	A	B
16	0	0	1	1	3	2	2	2	4	4	4	5	B	A	B
17	0	0	1	1	3	2	2	2	4	4	4	4	B	A	B
18	0	0	1	1	2	3	3	3	4	4	4	4	B	B	B
19	0	0	1	1	2	3	3	3	4	4	4	4	B	B	B
20	0	0	1	1	2	3	3	3	4	4	4	4	B	B	B

The design involves 60 exemplars with four features each, organized into three substructures. The 20 exemplars in the first substructure have features C_1-C_4, the 20 in the second substructure have features C_5-C_8 and the last 20 have features C_9-C_{12}. The feature has two possible values (0 vs. 1, for C_1-C_4; 2 vs. 3 for C_5-C_8; 4 vs. 5 for C_9-C_{12}). The last three columns headed by "Category" specify whether the exemplar is in category *A* or *B*. The first column is for exemplars with features C_1-C_4, the second for exemplars with features C_5-C_8, and the third for exemplars with features C_9-C_{12}.

In the first part of the experiment, the 60 exemplars are presented with feedback about whether they belong to category *A* or *B*. In the second part, the participants are asked to estimate the probability that probes with certain feature patterns belong to category *A*. There are *fifteen distinctive feature patterns*, six for features C_1-C_4, three for features C_5-C_8, and six for features C_9-C_{12}. The participants estimate the probability of category *A* for all fifteen patterns. The critical patterns are 1111 for features C_1-C_4, 3333 for C_5-C_8 and 5555 for C_9-C_{12}. Across these, the models provide distinctly different predictions (see Figure 1).

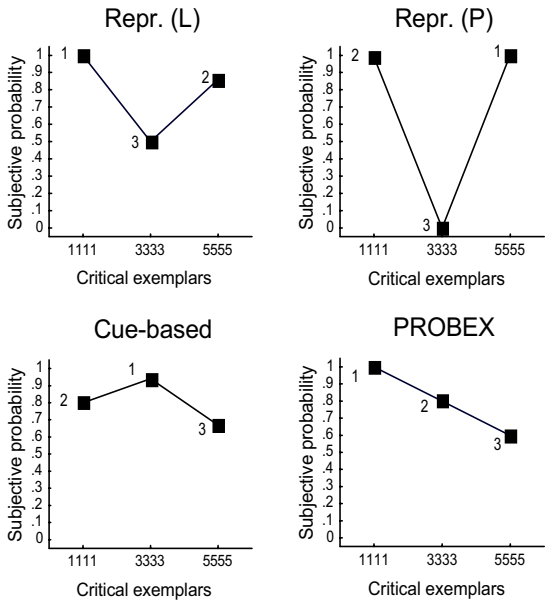


Figure 1: Predicted probability judgments. All predictions are derived with d=0. The predictions for representativeness with prototype similarity (P) are based on s=.1. The predictions for PROBEX are based s=0 (i.e., Picky frequentist).

For example, the predictions for feature pattern 3333 are derived as follows. The representativeness heuristic with a likelihood interpretation implies $p(A)$ = .25/(.25+.25) = .5: the probe is identical to 25% of the exemplars in category *A* and 25% of the exemplars in category *B*. In regard to a representativeness heuristic with prototype similarity, we note that the prototypes for category *A* and *B* in the second substructure (i.e., based on C_5-C_8) are 3222 and 2333, respectively. Ignoring the dampening *d*, Equation 2 implies the prediction $s^3/(s^3+s)$. The prototype for *A* differs on three features and the prototype for *B* on one feature. The prediction depends on the parameter *s*, but it will generally be low and always lower than .5. With cue-based relative frequency, $p(A)$ = 16/(16+1) = .94. Given the value of 3 for the most valid cue C_5, 16 of 17 exemplars belong to category *A*. According to the picky frequentist prediction by PROBEX (s=0), $p(A)$ = 4/5 = .8. Four out of five exemplars with identical feature patterns belong to

category *A*. At *s*>0, the prediction falls below .8. Predictions for the other two critical patterns are derived in the same way.

Note in Figure 1 that, depending on the model, the probability judgments for the three critical patterns have a different *rank order.* These predicted rank orders are *a priori* and not dependent on the parameters (i.e., *s* or *d*). By comparing the observed with the predicted rank order, we get a qualitative test of the models. In addition, we can evaluate the quantitative fit of the models to the judgments for all 15 feature patterns.

Method

Participants

Twenty-four undergraduate students (10 men and 14 women) in the age of 19 to 32 (average age = 23.3) participated. The participants where paid between 65-86 SEK depending on their performance. They received 30 SEK plus 1 SEK for each correct answer in the last learning block.

Apparatus and Materials

The experiment was carried out on a PC–compatible computer. In each of the four training blocks, the program first presented the 60 exemplars from Table 1. The task involved judgments for 60 companies, where 20 companies belonged to each of three countries (substructures). Each exemplar had four features that differed depending on the country. The features are presented in Table 2. The features and names of the countries were chosen to be as neutral as possible. In the test phase after each training block, the program presented each of the 15 distinct feature patterns twice.

Design and Procedure

A two-way within-subjects design was used. The independent variables were the number of training blocks (four blocks) and category substructure (three substructures). The dependent variable was the probability judgments. The specific assignment of concrete cue labels (see Table 2) to the abstract category structure (see Table 1) was varied and counterbalanced across the participants. Thus, each concrete label in Table 2 appeared equally often in each of the three substructures and equally often in the role of each of the abstract features denoted C_1 to C_{12} in Table 1.

The participants were to act as stockbrokers assigned to invest a large sum of money in three countries about which they knew nothing. They were told that it is usually enough to know four company features to know if the stock will rise or fall in the next twelve-months, but that the features differ between the countries.

Table 2: Twelve concrete features used in the experiment.

Features	Descriptions
1) Listed at the LAP / IPEK stock exchange?	
2) Less / more than 1000 employees?	
3) Commercials on television / the radio?	
4) Changed owner / merged in last three years?	
5) Less than / more than three years old?	
6) Give money to charity / sponsor sports team?	
7) Active in specific region / whole country?	
8) Co-operation with university / own research department?	
9) In state-financed SKATOS / TAPOS program?	
10) Primarily export-based / import-based?	
11) Affirmative action based on gender / ethnic background?	
12) Stock risen / fallen during the last 12-month?	

The participants were told that the first phase is a training session where they are presented with 60 companies, each described by four features that depend on the country. The features describe the companies as they were twelve months ago. They were to guess whether the stocks rose (*A*) or fell (*B*) in value in the last year. After each judgment, they received feedback on the actual development. The four features were presented on the screen. Below the question "Will the stock-value rise or fall during the next twelve month?" appeared. The participant answered s (short for the Swedish word for rise) or f (short for the Swedish word for fall). Thereafter, the correct answer appeared together with the company's four features.

In the test phase, the participants were told that they were to see a set of companies as they are today and judge the probability of an increase in their stock-value and that the markets are identical on all parameters today as they were one year ago. The feature patterns were presented in the same way as in the training phase, but with the question: "What is the probability that the stock of this company increases in value in the next 12 months?" They were told to answer in percentages and even up to 0, 10…100.

The test blocks consisted of two assessments of the 15 distinct feature patterns, one for rising stock-value (*A*) and one for falling stock-value (*B*). This allowed us to examine the additivity of the probability judgments (i.e., if the mean probability assigned to *A* and *B* for a feature pattern sum to 1). To get reliable data we recoded probability-*B* judgments into probability-*A* judgments by subtracting the probability-*B* judgments from 1. There was no feedback. The order of the probability judgments was counterbalanced within participants. The training and test blocks were repeated four times. The entire procedure took between one hour and fifteen minutes to two hours.

Results

Figure 2 presents mean probability judgments for the critical feature patterns in each of the four test blocks. The data for the third block shows a tendency to agree with the prediction by PROBEX. The fourth block exhibits clear agreement with the prediction by PROBEX. The confidence intervals for exemplars, 1111 and 5555 are clearly separated and the predicted decreasing trend is observed which refute all models except PROBEX.

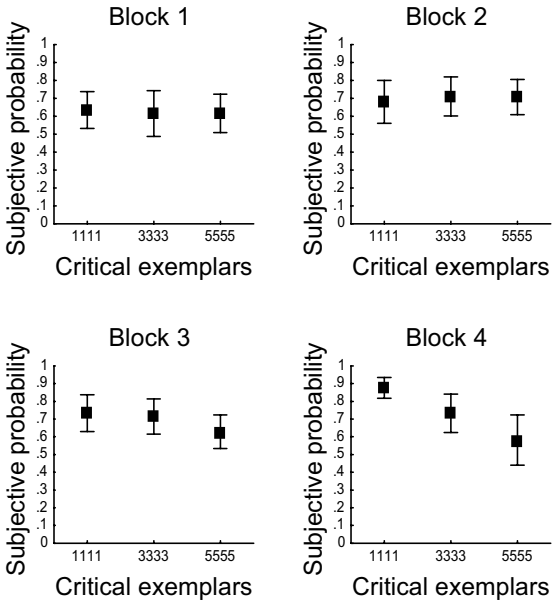


Figure 2: Means with 95% confidence intervals for the estimations of the critical exemplars for the four test blocks.

For the first two blocks, the data reveal no clear trend favoring any of the four models. One tentative interpretation of this result is that it reflects a mix of individual strategies in the early stages of training. To explore this more carefully, we fitted the four models to the data from all 15 distinct feature patterns. The probability judgments proved to be additive on average (i.e., the mean probability assigned to *A* and *B* for a feature pattern sum to 1).

The models were fitted to the mean probability judgments for each of the 15 distinct feature patterns with Root Mean Square Deviation (RMSD) as error function. This was done separately for each of the four test blocks. The model based on the representativeness heuristic as relative likelihood has one free parameter (*d*), representativeness heuristic as prototype similarity has two free parameters (*s* & *d*), cue-based relative frequency has one free parameter (*d*), and exemplar-based retrieval (PROBEX) has two free parameters (*s* & *d*). The results are summarized in Table 3.

Table 3 verifies that in the later stages of training, PROBEX provides a good fit to the data. Because the standard error of measurement is .05, the RMSDs for PROBEX (.054 & .058) come close to saturating the data. Considering all four blocks it is clear that cue-based relative frequency fits the judgments poorly in all blocks. Although the qualitative pattern in Figure 1 for blocks 1 and 2 does not accord with PROBEX, we find that it is the best fitting model throughout training. The models based on the representativeness heuristic exhibit moderate fit early in training, which successively deteriorates with training.

Table 3: Fit of the models as a function of test block in terms of RMSD and coefficients of determination r^2.

		Test Block			
Model	Index	1	2	3	4
Repr. (L)	RMSD	.087	.111	.105	.124
	r^2	.65	.69	.70	.73
Repr. (P)	RMSD	.094	.123	.124	.158
	r^2	.61	.62	.58	.55
Cue-based	RMSD	.139	.193	.188	.234
	r^2	.20	.21	.23	.22
PROBEX	RMSD	.060	.067	.054	.058
	r^2	.87	.92	.92	.95

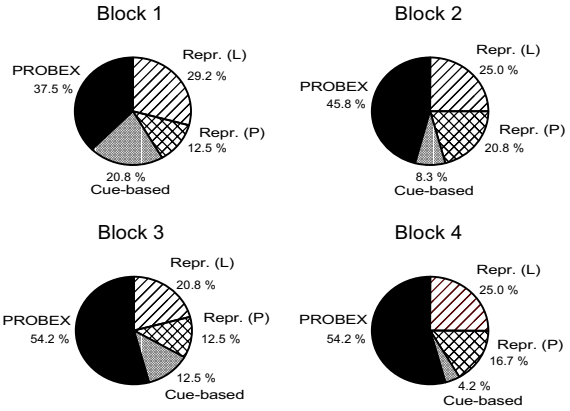


Figure 3: The percent of participants best described by each of the four models, in each of the four test blocks.

Finally, these conclusions were verified at the level of individual participants. The same model-fitting procedure was performed for each participant, with the exception that all models were fitted with one free parameter (*d*). In each block, the percentage of participants for which each model provided the best fit was ascertained. Figure 3 shows that PROBEX is the most frequent winner, although a minority of participants is

better fitted by representativeness as relative likelihood, mostly in the early test blocks.

Discussion

Research on subjective probability judgment has been characterized by a normative stance, where judgments are compared to norms from probability theory. Cognitive theory has primarily been evoked to provide post hoc explanations, as in most applications of the representativeness heuristic, or as scaffolds for more general predictions, as in the applications of cue-based relative frequency. The point of departure for our research is the need to make closer contact between cognitive theory and judgment research in controlled studies that allow us to support or refute core concepts in judgment research, such as the representativeness heuristic.

The results reported here provide clear support for the hypothesis of similarity-graded probability (Juslin & Persson, 2000). That an exemplar model is successful may not appear surprising considering the impressive performance of exemplar models in categorization studies (Nosofsky & Johansen, 2000). Yet, the results are at variance with crucial ideas in judgment research, like that of a representativeness heuristic (Kahneman et al., 1982) or cue-based relative frequency (Gigerenzer et al., 1991; Juslin, 1994).

The second to best fitting model was representativeness as relative likelihood, but this may be spurious as, the crucial feature patterns in Figure 1 aside, the predictions by the models tend to be correlated. However, the superiority of PROBEX is not a mere consequence of a greater inherent flexibility. To demonstrate this, we used the predictions for the last test block by representativeness as relative likelihood as fictive "true data" and added a normally distributed random error with a standard deviation of .05 to mimic measurement error. To this fictive data set, representativeness provided a superior fit (RMSD=.053, r^2=.97) as compared to PROBEX (RMSD=.096, r^2= .83). Thus, the better fit of PROBEX appears to reflect more than larger flexibility in the face of random error.

The best-fitting version of PROBEX (s=.21) in the last test block is not the Picky frequentist version identical to Bayesian estimation of the probability with a Beta-distribution (see Footnote 2). This suggests that, at least in regard to this more simplistic implementation of a Bayesian algorithm, PROBEX provides a better fit to data.

The main objection against the present study is perhaps that it is a single study involving one specific category structure. The category structure used here was guided by the aim of allowing qualitatively distinct predictions by the four models. This category structure may accidentally favor one model over another. Perhaps, a category structure more coherently organized around prototypes yields more support for representativeness as prototype similarity? Likewise, a more feature-rich category structure that posits more demand on information search may yield more support for cue-based relative frequency in the form of TTB (Gigerenzer et al., 1999). Only further research can tell. In any event, these hypotheses will have to count with a serious contestant in the form of PROBEX.

Acknowledgments

Bank of Sweden Tercentenary Foundation supported this research.

References

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101,* 519-527.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probalistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506-528.

Gigerenzer, G., Todd, P. M, & the ABC Research Group. (1999). *Simple heuristics that make us smart.* New York: Oxford University Press.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes, 57,* 226-246.

Juslin, P., & Persson, M. (2000). *Probabilities from exemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge*. Manuscript submitted for publication.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.) (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning, *Psychological Review, 85*, 207-238.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin and Review, 7,* 375-402.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 211-233.

Peterson, C. R, & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68*, 29-46.

Similarity Processing Depends on the Similarities Present: Effects of Relational Prominence in Similarity and Analogical Processing

Mark T. Keane (Mark.Keane@ucd.ie)
Department of Computer Science, Lukasiewicz Building,
University College Dublin, Belfield,, Dublin 4 IRELAND

Deirdre Hackett (Deirdre.Hackett@erc.ie)
Educational Research Center, Dromcondra, Dublin 9 IRELAND

Jodi Davenport (Jodi@mit.edu)
Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge, MA 02139-4307 USA

Abstract

Several studies have shown that similarity judgements involve a process of structural alignment akin to analogical mapping. Some research has shown that performing a similarity judgement task produces more relational responding in a subsequent cross-mapping task, suggesting that similarity necessarily uses structural alignment. However, other research has shown that this effect disappears when procedural/material manipulations fail to emphasise the relational aspects of similar scenes. The present study confirms the latter findings showing that if relational similarities are less prominent in a material set then subjects respond in an object-based rather than a relational way. Importantly, these results show that similarity processing does not by necessity make use of structural alignment but that the similarity processing adopted is pluralistic and depends on properties of the presented materials.

Introduction

A considerable body of recent research has shown that similarity comparisons can involve a process of structural alignment (see e.g., Goldstone, 1994; Goldstone & Medin, 1994; Goldstone, Medin & Gentner, 1991; Markman & Gentner 1993a, 1993b, 1997; Medin, Goldstone & Gentner, 1993). Representationally, this view characterises knowledge as structured hierarchies encoding objects, object attributes, relations between objects and relations between relations. Given these representations it is assumed that similarity comparisons involve the alignment of relational structure to find the most structurally consistent match between two systems of concepts, that satisfies the constraints of parallel connectivity (if two relations match, their arguments must match) and one-to-one mapping (that each item in one structure may only be mapped to one other item). Computationally, these ideas have been realised by a family of models that simulate analogical mapping (see e.g., Falkenhainer, Forbus & Gentner, 1989; Gentner, 1983, 1989; Holyoak & Thagard, 1989; Hummel & Holyoak, 1997; Keane, 1988, 1997; Keane & Brayshaw, 1988; Keane, Ledgeway & Duff, 1994; Veale & Keane, 1994, 1997, 1998). Indeed, structural alignment has been mooted as a unified account of a diverse range of phenomena including similarity, analogy, metaphor and concept combination (see Costello & Keane, 2000; Gentner, Holyoak & Kokinov, 2001; Keane & Costello, 2001).

Markman & Gentner (1993b) provided one of the key pieces of evidence supporting the role of structural alignment in similarity judgements. They used a one-shot mapping task in which subjects had to identify a cross-mapped object between two drawn scenes (see Appendix A). A cross-mapped object was defined as an object in one drawing that was perceptually similar to an object in a different relational role in the other drawing. So, for example, in the baseball scenes shown in Appendix A, the cross-mapped object would be the pitcher with the "C" on his uniform, because he is pitching in the upper scene and being pitched to in the other scene. Markman & Gentner have proposed that structural alignment is reflected in this task when subjects make relational responses (i.e., choosing the object in the same role) as opposed to object responses based on perceptual, feature similarity (i.e., choosing the perceptually similar object in a different role). The key manipulation asked participants to perform a similarity judgement task on the picture-pairs either before or after the mapping task. They found that

when participants made the similarity judgement *before* the mapping task they made more relational responses than when it was presented *after* the mapping task. Thus, the result strongly suggested that the similarity judgement task invoked a structural alignment process which then carried over to the mapping task increasing the proportion of relational responses (significantly, when an aesthetic-appreciation task was given before the mapping task no facilitation in relational responding was found).

Davenport & Keane (1999) queried these findings by pointing out that the materials and presentational procedure used may have contained unintended cues to promote relational responding. First, some of Markman & Gentner's materials had linguistic labels indicating the key relational similarity between the pictures (see e.g., the baseball pair in which "Pitch" is written). Second, the presentational procedure may have supported relational responding in that the all 8 stimuli presented were picture pairs with prominent relational similarities. Davenport & Keane found that when the linguistic cues were removed from the materials and the materials were mixed with fillers (involving simply object similarities) the similarity-judgement effect disappeared; that is, relational responding did not increase reliably when the similarity judgement task was made *before*, rather than *after* the mapping task. Significantly, Davenport & Keane also found that when the relational materials were blocked as a group and presented before the fillers there was increased relational responding relative to a condition in which the relational materials were randomly distributed among fillers in their order of presentation. This blocked-condition mimicked the presentational procedure used by Markman & Gentner. No reliable interaction was found between the ordering of the similarity judgement task (before or after mapping task) and the materials-order variable (blocked before or distributed among fillers). Hackett (2000) has subsequently replicated this materials order effect.

This pattern of results demands a very different account of similarity and mapping to that proposed by Markman & Gentner. First, it is no longer safe to assume that structural alignment is used in similarity judgements, as a matter of course, because relational responding does not follow prior similarity judgements. Second, it appears that relational responding is mainly dependent on the materials used and how they are presented. Specifically, the pattern of responding in the mapping task suggests that the *prominence of relational similarities* in the materials is the key variable affecting relational responding. This proposal best explains the evidence found:

- when linguistic cues are present that highlight key relational similarities then relational responding is seen (as in Markman & Gentner's original materials)
- when several materials are consecutively presented with relational similarities then relational responding results (as in Markman & Gentner's study and Davenport & Keane's blocked condition)
- when materials with clear object similarities are mixed up with these relational materials then relational responding decreases (as in Davenport & Keane's distributed condition)

So, structural alignment is really only used when the prominence of relational similarities in the materials appear to demand it, perhaps with the default style being processing based on attribute similarities.

If this account is true then any manipulation that reduces the prominence of relational similarities in the materials should reduce relational responding. For instance, if we take the materials that have previously produced relational responding and add in additional object similarities then, on balance, the relational similarities should be less prominent. Hence, we should see reduced relational responding for these modified materials. For example, consider picture-pair A in Appendix II; the top picture shows a woman kicking a football with goal posts and a sun behind her and the bottom picture shows her being kicked by a child with some blocks and a clock in the background. This picture pair is quite sparse, a sparseness that lends a greater prominence to the kicking relation shown. Compare picture-pair A with picture-pair B in Appendix II; the latter also shows two kicking episodes but the scenes are much richer with more similar objects in the top and bottom pictures (houses, the sun, goal posts, etc). Although, both picture pairs have the same kicking incident, the greater frequency of object similarities in the richer pair should, if our hypothesis is correct, reduce the occurrence of relational responding relative to the sparser pair.

The present study examines this sparse versus rich manipulation, where the richer pairs were essentially the same scenes with added matching objects. We also attempted a further replication of the similarity-judgement effect by giving different groups a similarity judgement task either before or after the mapping task. As such, we had a 2 x 2 between-subject design where the variables were task-order (similarity task before or after mapping task) and material-type (sparse or rich materials). Following Davenport & Keane, all conditions presented the materials in a distributed fashion with the target materials being randomly distributed with fillers. We made two predictions in the study. First, that the similarity judgement task would have no effect on relational responding, confirming Davenport & Keane's finding. Second, that the sparse materials would produce more relational responding than the rich materials, as the prominence of the relational

similarities is reduced in the latter by the greater frequency of object similarities.

Method

Subjects. Forty-eight undergraduate students at University College Dublin took part voluntarily in the experiment and were randomly assigned to one of the four between-subjects conditions.

Stimuli. The stimuli for this experiment consisted of 8 pairs of pictures depicting causal scenes with matching relational structure and 16 filler pairs. Each of these 8 pairs contained a cross-mapping as operationalized by Markman and Gentner (1993b) in which a pair of perceptually-similar objects were shown which played different roles in the matching relational structure of the two scenes (see Appendix B for an example). In all four conditions, the 8 relational pictures were designed so that the perceptually-similar target object was in approximately the same spatial position in the picture pairs (e.g., the woman in the soccer materials is in the same central position in both pictures). Two versions of the relational materials were prepared. The rich set was created by adding similar objects to both scenes of the original sparse pictures used by Davenport & Keane (see Appendix II for an example). In all other respects, the picture-pairs were the same (e.g., in the placing of the arrow indicating the to-be-mapped object).

Eight of the filler pairs depicted comparable scenarios without matching relational structure (e.g. two beach scenes, one with a man surfing another with a child is building a sand castle) and the other 8 pairs did not match in either scenario or relational structure (e.g. a scene of an artist and a scene of a man in a grocery store). The fillers were the same as those used by Davenport & Keane.

The stimuli were presented in booklet form with one pair on each page (one picture above the other). The stimuli for the mapping task had an arrow placed above an object in the top scene. For the 8 target pairs this was the cross-mapped object, otherwise it was an object which appeared in both scenes. The stimuli used for the similarity rating task had a scale with the numbers 1 through 9 at the bottom of the page. The words Low Similarity appeared under the 1 and the words High Similarity appeared under the 9.

Booklets in the all conditions had a completely randomised presentation of the 24 pairs for both the mapping and the similarity tasks.

Procedure. As in Gentner & Markman's study, the first page of the mapping section of the booklet instructed subjects to draw a line from the object under the arrow to the object in the bottom scene that "best went with that object". The first page of the similarity judgement section instructed subjects to rate the similarity of the two scenes by circling a number on the scale at the bottom of the page.

Subjects in the similarity-after conditions received a booklet with the mapping task followed by the similarity judgement task while subjects in the similarity-before condition received a booklet with the similarity judgement task first.

Subjects were tested in small groups of varying sizes and each experimental session took between 10 and 15 minutes.

Scoring. As in Gentner & Markman's study, participants' responses to the 8 target pairs in the mapping task were determined as an *object mapping* if a line was drawn from the cross-mapped object to a featurally-similar object in the bottom scene; a *relational mapping*, if a line was drawn to the object in the same relational role in the bottom scene; or a *spurious mapping* if a line was drawn to another, unrelated object. As in previous studies, spurious responses were removed prior to data analysis (5% of all responses made).

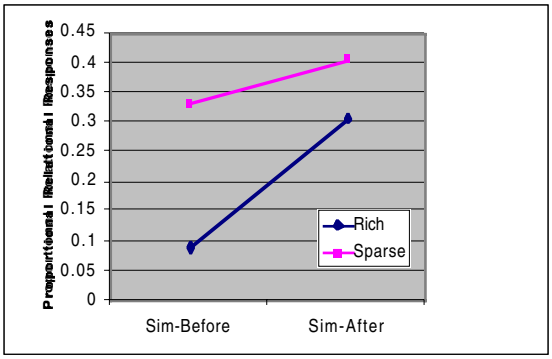


Figure 1: The Mean Proportion of Relational Responses in the Conditions of the Study

Results & Discussion

A two-way, between-subjects ANOVA found a reliable effect of material-type with a higher proportion of relational responding in the sparse conditions (M = .37) than in the rich conditions (M=0.19), though this difference is marginally reliable, $F(1,44) = 3.538$, p=.06, $MSError = 0.079$ (see Figure 1). Again, contrary to Markman and Gentner's predictions, no reliable effect was found for the task-order variable, $F(1,44) = 2.279$, $p > .10$. Indeed, the direction of the difference is in the opposite direction to that predicted; *less* relational responding occurs when the similarity judgement task is *before* as opposed to *after* the mapping task. Finally, there was no reliable interaction between task-order and material-type, $F(1,44) = 0.804$, $p > .10$.

General Discussion

There are two significant findings in this study. First, it replicates Davenport & Keane's finding that

relational responding is not influenced by a prior similarity judgement task; leaving open the question of whether similarity judgement involves structural alignment. Indeed, if we adopt the argument made by Markman & Gentner, we would be bound to conclude that structural alignment does not necessarily occur in similarity judgements. Second, and perhaps more surprisingly, it shows that structural alignment does not necessarily occur in the mapping task either. Rather, it appears that people sometimes respond on the basis of relational similarities and other times respond on the basis of object similarities. Furthermore, the key factor determining the mode of response lies in the nature of the materials themselves. When relational similarities are prominent in the given materials or across a set of consecutive materials then relational responding will result, but when these relational similarities are counter balanced by more object similarities in a given material or across a set of consecutive materials then object-based responding results.

Computationally, these findings present a number of challenges for models of similarity and analogy. They suggest that there are two distinct modes of processing for similarity judgement and mapping tasks. In one mode, relational correspondences are mainly used; this could be achieved by only computing relational matches (e.g., a type of selective attention to relations) or by computing both object and relational matches and then subsequently giving a greater weight to relational similarities. In the other mode, object attributes are mainly used: this could be achieved by only computing attribute matches (e.g., a type of selective attention to object features) or by computing both object and relational matches and then subsequently giving a greater weight to object similarities. Where you have two modes of processing there has to be a trigger for switching processing from one mode to the other. The empirical evidence suggests that this trigger is sensitive to the relative frequency of relational versus attribute similarities present in a stimulus pair and a set of stimulus pairs.

Do any current models have these sort of properties ? Goldstone's (1994) interactive activation model gave a greater weight to attribute matches during early stages of processing with relational matches emerging later on; this model deals with the finding that under time pressures people rely more on attribute similarity (Medin & Goldstone, 1994). However, it is not immediately clear whether it would predict less relational responding in the rich versus the sparse materials. Furthermore, this model would need to have a history of previous similarity judgements to model the effects of consecutive relational materials. Another model, the MAX model (Goldstone, Medin & Gentner, 1991) pools relational and attribute similarities separately, with relational or attribute responses being chosen based on the relative sizes of the two pools of similarities. This model might be able to deal with the relational prominence effect, as the rich materials might have a larger pool of attribute similarities relative to the sparse materials, leading to an object-based response. However, the MAX model does not have an account of the effects of consecutive materials as, again, it holds no history of previous similarity episodes. In short, no current model seems to be able to handle this evidence.

The current results leave open the question of whether similarities are selectively attended to with only a subset of all possible similarities being computed or whether all similarities are computed and then evaluated to decide on what response mode should be adopted. This is a key question to be settled by future empirical work. They also leave open the question of which mode of processing is the default mode or, in indeed, if there is a default mode. Attribute similarities seem to be computed more efficiently and easily (Goldstone & Medin, 1994). This mode of processing also has a laziness that is characteristic of human cognitive processing, when one considers people's mental sloth in making elaborative, bridging relational inferences.

However, there is one conclusion that is unavoidable given the current results: Namely, that comparison processes -– whether they be similarity or analogical processes – are pluralistic rather than monolitic (see Medin, Goldstone & Gentner, 1994; Keane & Costello, 2001). This conclusion should shift the focus of research to the key issue of what variables influence the mode of processing adopted in this pluralistic computational environment.

References

Costello, F. & Keane, M.T. (2000). Efficient creativity: Constraints on conceptual combination.*Cognitive Science.*

Davenport, J. & Keane, M.T. (1999). Similarity & structural alignment: You can have one without the other. In *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Falkenhainer, B., Forbus, K.D., & Gentner, D. (1989). Structure-mapping engine. *Artificial Intelligence, 41,* 1-63.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7* , 155-170.

Gentner, D. (1989). Mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. Cambridge: CUP.

Gentner, D., Holyoak, K.J., & Kokinov, B. (2001). *The Analogical Mind.* Cambridge, MASS; MIT Press.

Goldstone, R.L. (1994). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Memory and Cognition,* 20, 3-28.

Goldstone, R.L. & Medin, D.L. (1994). Time course of comparison. *Journal of Experimental Psychology: Language, Memory & Cognition*, 20, 29-50.

Goldstone, R.L., Medin, D.L., & Gentner, D. (1991). Relational similarity and the non-independence of features in similarity judgments. *Cognitive Psychology, 23,* 222-262.

Hackett, D. (2000). *Similarity judgements and the causes of relational responding.* MSc Thesis, Department of Computer Science, University College Dublin, Dublin, Ireland.

Holyoak, K.J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science, 13,* 295-355.

Hummel, J.E. & Holyoak, K.J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review. 104,* 427-466.

Keane, M.T. (1988). *Analogical Problem Solving.* Chichester, England: Ellis Horwood (Cognitive Science Series). [Simon & Schuster in N. America]

Keane, M.T. (1997). What makes an analogy difficult ?: The effects of order and causal structure in analogical mapping. *Journal of Experimental Psychology: Language, Memory & Cognition, 23,* 946-967.

Keane, M.T. & Costello, F. (2001). Why Conceptual Combination is Seldom Analogy. In D. Gentner, K.J. Holyoak, & B. Kokinov (Eds.), *The Analogical Mind.* Cambridge, MASS; MIT Press.

Keane, M.T., & Brayshaw, M. (1988). The Incremental Analogical Machine: A computational model of analogy. In D. Sleeman (Ed.), *European Working Session on Machine Learning.* London: Pitman.

Keane, M.T., Ledgeway, T. & Duff, S. (1994). Constraints on analogical mapping: A comparison of three models. *Cognitive Science*, 18, 387-438.

Markman, A.B. & Gentner, D. (1993a). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language, 32,* 517-535.

Markman, A.B. & Gentner, D. (1993b). Structural alignment during similarity comparisons. *Cognitive Psychology, 25,* 431-467.

Markman, A.B. & Gentner, D. (1997). The effects of alignability on memory. *Psychological Science, 5,* 363-367.

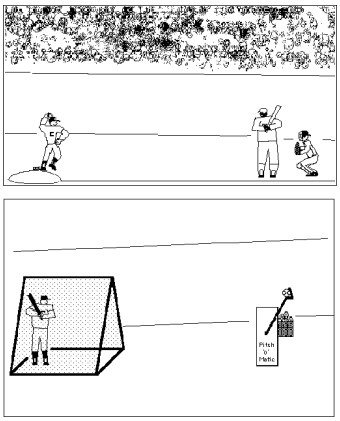
Medin, D.L., Goldstone, R.L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100,* 254-278.

Veale, T. & Keane, M.T. (1994). Belief modelling, intentionality and perlocution in metaphor comprehension. *Proceedings of the Sixteenth Annual Meeting of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Veale, T. & Keane, M.T. (1997). The competence of sub-optimal structure mapping on 'hard' analogies. *IJCAI'97: The 15th International Joint Conference on Artificial Intelligence.* Morgan Kaufmann.

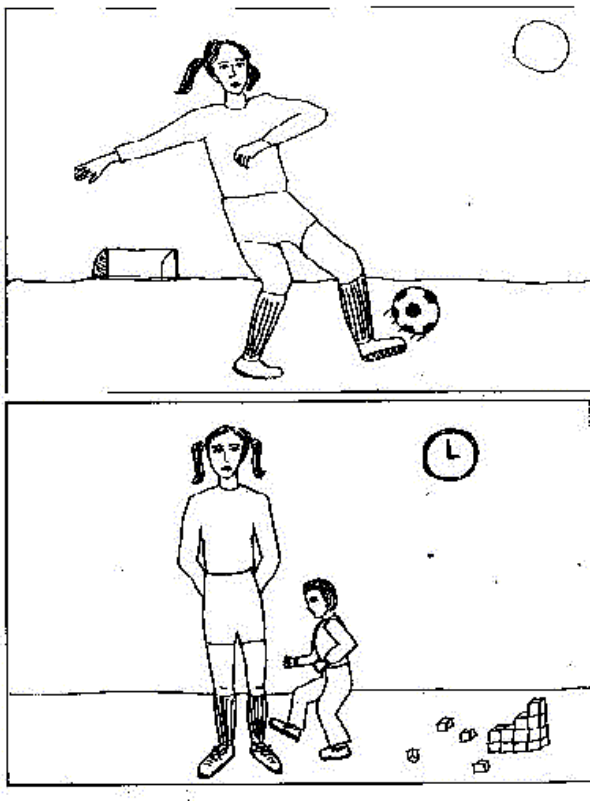
Veale, T. & Keane, M.T. (1998). Principle Differences in Structure Mapping. In K.J. Holyoak, D. Gentner, & B. Kokinov (Eds.), *Proceedings of Analogy '98.* New University of Bulgaria Press: Bulgaria.

Appendix I. The Baseball Materials Used by Markman & Gentner(1993b)



Appendix II. Examples of the materials used in the study showing the (a) spare kick material and (b) the rich kick material.

A.



B.

Constraints on Linguistic Coreference: Structural vs. Pragmatic Factors

Frank Keller
keller@coli.uni-sb.de
Computational Linguistics, Saarland University
PO Box 15 11 50, Building 17.1
66041 Saarbrücken, Germany

Ash Asudeh
asudeh@csli.stanford.edu
Department of Linguistics, Stanford University
Margaret Jacks Hall, Building 460
Stanford CA 94305-2150, USA

Abstract

Binding theory is the component of grammar that regulates the interpretation of noun phrases. Certain syntactic configurations involving picture noun phrases (PNPs) are problematic for the standard formulation of binding theory, which has prompted competing proposals for revisions of the theory. Some authors have proposed an account based on structural constraints, while others have argued that anaphors in PNPs are exempt from binding theory, but subject to pragmatic restrictions. In this paper, we present an experimental study that aims to resolve this dispute. The results show that structural factors govern the binding possibilities in PNPs, while pragmatic factors play only a limited role. However, the structural factors identified differ from the ones standardly assumed.

Introduction

Linguistic Intuitions The data on which linguists base their theories typically consist of grammaticality judgments, i.e., intuitive judgments of the well-formedness of utterances in a given language. When a linguist obtains a grammaticality judgment, he or she performs a small experiment on a native speaker; the resulting data are behavioral data in the same way as other measurements of linguistic performance (e.g., the reaction time data used in psycholinguistics). However, in contrast to experimental psychologists, linguists are generally not concerned with methodological issues, and typically none of the standard experimental controls are imposed in collecting data for linguistic theory. As Schütze's (1996) recent work on empirical issues in linguistics demonstrates, such methodological negligence can seriously compromise the data obtained. Schütze (1996) argues for a more reliable mode of data elicitation in linguistics, based on standard methods from experimental psychology.

Recently, Bard, Robertson, and Sorace (1996) and Cowart (1997) demonstrated how the experimental paradigm of magnitude estimation (ME) makes it possible to address problems such as the ones raised by Schütze. ME is an experimental technique standardly used in psychophysics to measure judgments of sensory stimuli (Stevens, 1975). It requires subjects to estimate the magnitude of physical stimuli by assigning numerical values proportional to the stimulus magnitude they perceive. Highly stable judgments can be achieved for a whole range of sensory modalities, such as brightness, loudness, or tactile stimulation. Bard et al. (1996) demonstrated that linguistic judgments can be elicited in the same way as judgments of sensory stimuli, and that ME can yield reliable and fine-grained measurements of linguistic intuitions.

The present paper applies the ME methodology to a longstanding dispute in linguistic theory, viz., the binding theoretic status of picture noun phrases (PNPs). Binding in PNPs has generated considerable interest in the literature, and has prompted a number of revisions of standard binding theory. However, there is considerable disagreement on both the relevant data (i.e., coreference judgments for PNPs) and on the theoretical conclusions to be drawn from these data. In this paper, we demonstrate how the use of experimentally elicited coreference judgments can resolve such theoretical disputes.

Binding Theory and Exempt Anaphors Binding theory (BT) is the component of grammar that regulates the interpretation of noun phrases (NPs). Three types of noun phrases are generally distinguished: (a) full NPs (e.g., *Hanna*, *the woman*, *every woman*), (b) pronouns (e.g., *he*, *her*), and (c) anaphors (e.g., *herself*, *each other*). The task of BT is to determine which noun phrases in a given syntactic domain can be *coreferential*, i.e., refer to the same individual. Coreference is normally indicated with subscripts:

(1) a. Hanna$_i$ admires *her$_i$/herself$_i$.
b. Hanna$_i$ thinks that Peter admires her$_i$/*herself$_i$.

In example (1a), the proper name *Hanna* and the pronoun *her* cannot refer to the same person, i.e., they cannot be coreferential (as indicated by the '*'). The pronoun cannot be *bound* by the proper name. In (1b), on the other hand, *Hanna* is a potential binder for *her*, i.e., coreference is possible. The possibilities for anaphor binding are exactly reversed; *Hanna* must bind (i.e., corefer with) *herself* in (1a), but cannot do so in (1b).

There are distinct structural conditions that determine the binding possibilities for the different kinds of NPs. Principle C of BT deals with the binding requirement for full NPs, and will not concern us here. Principle A captures the binding requirements for anaphors; in early formulations, it states that an anaphor has to be bound within a certain local domain (Chomsky, 1981). The local domain is defined using c-command, a structural notion defined on trees. Principle B, on the other hand, states that pronouns cannot be bound within the local domain. It follows that anaphors and pronouns are predicted to be in complementary distribution, i.e., anaphors can be bound where pronouns cannot be bound, and vice versa.

It was subsequently observed that this complementarity breaks down in certain structures. A case that has generated much theoretical discussion is PNPs, where anaphors and pronouns are equally acceptable:

(2) Hanna$_i$ found a picture of her$_i$/herself$_i$.

There is also the further complication that in PNPs with possessors (3) and in PNPs that are arguments of certain verbs (4) the complementarity between pronouns and anaphors seems to resurface:

(3) Peter$_i$ found Hanna$_j$'s picture of *her$_j$/herself$_j$.
(4) Hanna$_i$ took a picture of *her$_i$/herself$_i$.

Note that (4) is meant with the sense of *take* as in creating a photograph, not as in physically removing a picture. A number of authors have argued for a revised version of BT based on data such as (2), (3), and (4). Chomsky (1986) restates BT such that there is an asymmetry between pronouns and anaphors in certain contexts, including PNPs without possessors. For (4), Chomsky and Lasnik (1995) propose that there is a covert possessor. With these revisions, the predicted pattern of data is exactly as in (2)–(4). We will refer to this approach as the *structural account* of binding in PNPs.

Some more recent work, however, has proposed a *pragmatic* account of the PNP data in (2)–(4) (e.g., Kuno, 1987; Pollard & Sag, 1994; Reinhart & Reuland, 1993). These authors have observed that in certain configurations anaphors are *exempt* from BT. One such configuration is PNPs without possessors, as in (2) and (4). According to this view, the anaphor in (2) is not subject to Principle A, but is rather governed by pragmatic constraints, where relevant factors include referentiality, definiteness, and aspect. It is important to note that even the versions of BT that postulate exempt anaphors still maintain that Principle A holds of anaphors in PNPs when there is an overt possessor: although the anaphor in (2) is exempt, the anaphors in (3) and (4) are still subject to BT.

The present study attempts to clarify the empirical status of exempt anaphors. We present the results of an experiment that tests the influence of both structural and pragmatic factors on coreference in PNPs. This experiment uses the magnitude estimation (ME) paradigm to establish the coreference intuitions of linguistically naive subjects. (For other studies demonstrating the usefulness of experimental data in clarifying BT facts, see Cowart, 1997; Gordon & Hendrick, 1997.)

Before we discuss the results of this experiment, we present a control study designed to validate our experimental paradigm. To our knowledge, ME has never been applied to coreference judgments, hence we must show that its results are consistent with the theoretical literature and replicate previous experimental data.

Experiment 1: Control Study

The control study was designed as a replication of Experiment 3 of the study of coreference by Gordon and Hendrick (1997). It investigated basic effects of Principles A, B, and C of BT. Eight different binding configurations were tested, three of which occurred either with or without c-command (see Chomsky, 1981, for details on c-command). Table 1 lists the binding configuration tested by Gordon and Hendrick (1997). It also summarizes the predictions of standard BT for these configurations, and gives example stimuli.

Predictions

Our hypothesis is that ME generates valid coreference judgments. We therefore predict that the same significant effects as in Gordon and Hendrick's (1997) original study will be present, even though our replication used an ME task instead of the ordinal judgment task employed by Gordon and Hendrick (1997). Another difference is that we conduced our experiment over the World Wide Web, while the Gordon and Hendrick (1997) administered a conventional questionnaire. The web-based methodology entails differences in sampling and experimental procedure, which increases the need for a validation study.

Method

Subjects Fifteen participants were recruited over the Internet by postings to newsgroups and mailing lists. All participants were self-reported native speakers of English and naive to syntactic theory.

Materials Following Gordon and Hendrick (1997), the design contained one factor, viz., binding configuration (*Ana*) with eight levels. Three lexicalizations were used; one was the original lexicalization used by Gordon and Hendrick (1997), the other two were new lexicalizations, designed in analogy with the original one. This resulted in a set of 24 items (see Table 1 for sample stimuli).

Procedure The method used was ME as proposed by Stevens (1975) for psychophysics and extended to linguistic stimuli by Bard et al. (1996) and Cowart (1997).

Subjects first saw a set of instructions that explained the concept of numerical ME using line length. Subjects were instructed to make length estimates relative to the first line they would see, the reference line. They were told to give the reference line an arbitrary number, and then assign a number to each following line so that it represented how long the line was in proportion to the reference line. Several example lines and corresponding numerical estimates were provided to illustrate the concept of proportionality. Then subjects were told that linguistic acceptability could be judged in the same way as line length, and that this experiment required them to judge the acceptability of coreference. Following Gordon and Hendrick (1997), this was defined as follows: 'Your task is to judge how acceptable each sentence is by assigning a number to it. By acceptability we mean the following: Every sentence will contain two expressions in ALL CAPITALS. A sentence is acceptable if these two expressions can refer to the same person.' The task was illustrated by examples.

After reading the instructions, subjects took part in a training phase designed to familiarize them with the task. In the training phase, subjects were ask to use ME to judge the length of a set of lines. Then, a set of practice items (similar to the experimental items) were administered to familiarize subjects with applying ME to linguistic stimuli. Finally, subjects had to judge the experimental items. Each subject judged all 24 experimental stimuli and a set of 24 fillers, i.e., a total of 48 items.

The experiment was conducted over the web using WebExp 2.1 (Keller, Corley, Corley, Konieczny, & Todirascu, 1998), an interactive software package for web-based psycholinguistic experimentation. Keller and Alexopoulou (2001) present a detailed discussion of the safeguards that WebExp puts in place to ensure the authenticity and validity of the data collected, and also present a validation study comparing web-based and lab-based judgment data (for a WebExp validation study using sentence completion data, see Corley & Scheepers, in press).

Results

The data were normalized by dividing each numeric judgment by the modulus value that the subject had assigned to the reference sentence. This operation creates a common scale for all subjects. Then the data were transformed by taking the decadic logarithm. This transformation ensures that the judgments are normally distributed and is standard practice for ME data (Bard et al., 1996). All analyses and figures are based on normalized, log-transformed judgments.

The average judgments for the different conditions are graphed in Figure 1 for the original study and in Figure 2 for our replication. Visual inspection shows that the replication experiment produces the same acceptability pattern for each of the binding configurations.

This was confirmed by the statistical analyses. Gordon and Hendrick (1997) report a significant main effect of binding configuration (*Ana*), which was also present in our data ($F_1(7,98) = 17.561$, $p < 0005$; $F_2(7,14) = 295.262$, $p < 0005$). They also found that the acceptability of the name-anaphor configuration increased under c-command, which was replicated in our data ($F_1(1,14) = 17.057$, $p = .001$; $F_2(1,2) = 2389.474$, $p < 0005$). Another finding was that c-command significantly reduces the acceptability of coreference name-pronoun configurations. This effect was also present in the replica-

Table 1: Sample stimuli and predictions from Gordon and Hendrick (1997), Experiment 3

NP_1	NP_2	c-command	sample sentence	prediction
name	pronoun	no	(i) **Joan's** father respects **her**.	grammatical
pronoun	name	no	(ii) **Her** father respects **Joan**.	grammatical
name	name	no	(iii) **Joan's** father respects **Joan**.	grammatical
pronoun	anaphor	no	(iv) **Her** father respects **herself**.	ungrammatical
name	anaphor	no	(v) **Joan's** father respects **herself**.	ungrammatical
name	pronoun	yes	(vi) **Joan** respects **her**.	ungrammatical
pronoun	name	yes	(vii) **She** respects **Joan**.	ungrammatical
name	anaphor	yes	(viii) **Joan** respects **herself**.	grammatical

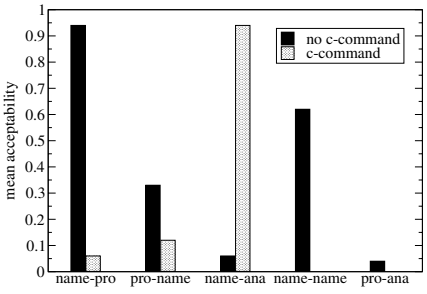


Figure 1: Original data from Gordon and Hendrick (1997), Experiment 3

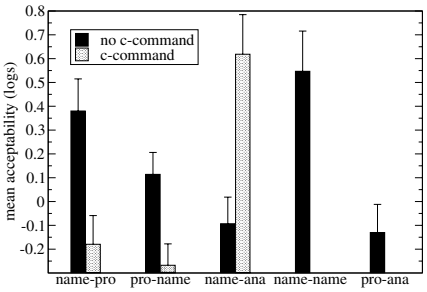


Figure 2: Replication of Gordon and Hendrick (1997), Experiment 3

tion ($F_1(1,14) = 21.818$, $p < .0005$; $F_2(1,2) = 315.306$, $p = .003$). An effect of c-command on the acceptability of pronoun-name configurations was also found both in the original data set and in our replication ($F_1(1,14) = 25.949$, $p < .0005$; $F_2(1,2) = 181.980$, $p = .005$).[1] Finally, a comparison of the name-pronoun and the name-name configurations showed that names are favored as antecedents ($F_1(1,14) = 13.770$, $p < .002$; $F_2(1,2) = 192.301$, $p = .005$), in line with what Gordon and Hendrick (1997) found.

To further compare the results of the original experiment and our validation study, we conducted a correlation analysis comparing the mean judgments for each cell in the experiment. A high correlation coefficient was obtained ($r_1 = .9198$, $p = .001$, $N = 8$). (No by-item correlation coefficient could be computed as Gordon and Hendrick (1997) fail to report by-item analyses.)

[1]Note that standard BT fails to predict the reduced acceptability of configuration (ii). A possible explanation might be that this configuration involves cataphoric reference (i.e., the pronoun refers forwards instead of backwards).

Discussion

In this study, we used ME to replicate a published experiment on coreference judgments that used a conventional ordinal scale (Gordon and Hendrick's (1997) Experiment 3). We obtained the same significant effects as in the original and a high correlation with the original data set, which amounts to a full replication of the original study.

This result indicates that the ME paradigm is suitable for investigating judgments of linguistic coreference, which are vital for testing claims from BT. Previous uses of ME were limited to grammaticality judgments (Bard et al., 1996; Cowart, 1997). The successful replication also reassures us that psycholinguistic data collected over the web yield results comparable to data generated by a conventional lab-based methodology, in line with previous findings by Keller and Alexopoulou (2001) and Corley and Scheepers (in press).

Finally, the present experiment allows us to establish a baseline for further experiments on linguistic coreference. It encompassed only clear-cut cases of coreference that are uncontroversial in the binding theoretical literature. It is important to establish the validity of our methodology for such clear-cut cases before moving to investigate more controversial issues such as binding in PNPs, where the theoretical and empirical claims in the syntactic literature differ widely. Binding in PNPs is the subject of the next experiment.

Experiment 2: Structural and Pragmatic Factors in Coreference

Based on the results from the control experiment, we carried out an experimental study investigating the factors that determine coreference in PNPs in English. The aim of this experiment was to provide reliable experimental data that settles the longstanding dispute about the binding theoretical status of PNPs. In particular, we tested the claim that PNPs are exempt from BT, and hence their coreference options are governed by pragmatic, rather than structural factors.

Structural Factors The current experiment tested the influence of structural factors on binding in PNPs by comparing the behavior of anaphors and pronouns in six configurations, listed in Table 2. Two structural factors were tested.

Firstly, the position of the binder, which can either be the subject of the matrix clause (as in configurations (i)–(iv) in Table 2), or the possessor of the PNP (as in configurations (v) and (vi) in Table 2). Secondly, the absence of a possessor (as in configurations (i) and (ii)), or its presence (as in configurations (iii)–(vi) in Table 2). The experiment contained three subdesigns, which tested the configurations (i) and (ii), (iii) and (iv), and (v) and (vi), respectively.

Table 2: Sample stimuli and predictions for Experiment 2

NP_1	NP_2	subject	possessor	sample sentence	prediction
name	pronoun	yes	no	(i) **Hanna** found a picture of **her**.	grammatical
name	anaphor	yes	no	(ii) **Hanna** found a picture of **herself**.	grammatical
name	pronoun	yes	yes	(iii) **Hanna** found Peter's picture of **her**.	grammatical
name	anaphor	yes	yes	(iv) **Hanna** found Peter's picture of **herself**.	ungrammatical
name	pronoun	no	yes	(v) Hanna found **Peter's** picture of **him**.	ungrammatical
name	anaphor	no	yes	(vi) Hanna found **Peter's** picture of **himself**.	grammatical

Pragmatic Factors The second aim of the present experiment was to investigate the influence of pragmatic factors on the coreference in PNPs. Such factors have received much attention in the theoretical literature. However, no quantitative studies have been conducted to determine to what extent these factors influence coreference, and how they interact with structural factors.

Three pragmatic factors were investigated. The first one is definiteness of the PNP. As an example of definiteness consider the minimal pair in (5): the PNP in (5a) is indefinite and the one in (5b) is definite.

(5) a. Hanna$_i$ found a picture of her$_i$/herself$_i$.
b. Hanna$_i$ found the picture of her$_i$/herself$_i$.

The second factor is the aspectual class of the matrix verb, illustrated in example (6): *find* and *lose* are examples of achievement verbs, while *take* and *destroy* are accomplishment verbs; *find* and *take* are [existence], i.e., they presuppose the existence of their object, while *lose* and *destroy* are [−existence], i.e., they do not carry this presupposition.

(6) a. Hanna$_i$ found a picture of her$_i$/herself$_i$.
b. Hanna$_i$ lost a picture of her$_i$/herself$_i$.
c. Hanna$_i$ took a picture of her$_i$/herself$_i$.
d. Hanna$_i$ destroyed a picture of her$_i$/herself$_i$.

Third, we tested the influence of the referentiality of the binder, as illustrated in (7):

(7) a. Hanna$_i$ found Peter's picture of her$_i$/herself$_i$.
b. The woman$_i$ found Peter's picture of her$_i$/herself$_i$.
c. Each woman$_i$ found Peter's picture of her$_i$/herself$_i$.

The pragmatic factors were included in the three subdesigns of the present experiment. The factors definiteness and verb class were included in the first subdesign, while referentiality was part of the second and third subdesign.

Predictions

Based on the theoretical literature (see Introduction), we predict that anaphors in PNPs are exempt from local binding (i.e., binding within the PNP), *unless* the PNP has a possessor, in which case the anaphor must be bound by the possessor (see examples (2) and (3)). We also predict that pronouns must be locally free from a possessor, if there is one. Table 2 lists the configurations and the associated predictions, together with example stimuli. Note that we expect that the relative acceptability of pronouns and anaphors is the same in configurations (i), (ii) and (iii). Configurations (iv) and (v) are predicted to be unacceptable, while (vi) is predicted to be acceptable. These constructions differ in terms of their syntactic structure (antecedent is the subject or the possessor; possessor is present or not). We expect to find no main effect of binding configuration for (i) versus (ii), but for pairs (iii)/(iv) and (v)/(vi) we expect binding configuration to have a significant main effect.

If the pragmatic approach to binding in PNPs is correct, then we also expect that the pragmatic factors verb class, definiteness, and referentiality have an effect on coreference. The underlying theoretical assumption is that coreference for exempt NPs is governed by pragmatics, rather than by structural principles. Hence we predict an interaction of binding configuration with verb class and definiteness in the first subexperiment, and an interaction of binding configuration with referentiality in the second and third subexperiment.

Method

Subjects Fifty-two native speakers of English volunteered to participate. All participants were naive to syntactic theory.

Materials The experimental materials included three subdesigns. The first subdesign investigated binding configurations (i) and (ii): name-pronoun and name-anaphor with the antecedent in the subject and without a possessor. The second subdesign compared binding configurations (iii) and (iv): name-pronoun and name-anaphor with the antecedent in the subject and a possessor in the PNP. The third subdesign dealt with configurations (v) and (vi): name-pronoun and name-anaphor with the antecedent as the possessor of the PNP.

This means that in each of the three subdesigns the factor binding configuration (*Ana*) had two levels: name-pronoun or name-anaphor. In the first subdesign, this factor was crossed with *Def* and *Verb*, representing the two pragmatic factors definiteness of the PNP and aspectual class of the main verb. *Def* had two levels (definite, indefinite, see (5)), *Verb* had three levels (achievement [existence], accomplishment [existence], accomplishment [−existence] (see (6a), (6c), (6d)). This yielded a total of $Ana \times Def \times Verb = 2 \times 2 \times 3 = 12$ cells for the first subdesign.

In the second and third subdesigns, the structural factor *Ana* was crossed with the pragmatic factor referentiality (*Ref*), which had three levels (proper name, definite NP, quantified NP, see (7)). The second and third subdesign therefore had $Ana \times Ref = 2 \times 3 = 6$ cells each.

All three subdesigns taken together had a total of 24 cells. Four lexicalizations were used for each of the cells, which resulted in a total of 96 stimuli. A set of 24 fillers was used, designed to cover the whole acceptability range.

Procedure The experimental procedure was the same as in Experiment 1. The stimulus set was divided into four subsets of 24 stimuli by placing the items in a Latin square. Each subject judged one of these subsets and 24 fillers, i.e., a total of 48 items.

Results

The data were preprocessed as in Experiment 1. Separate ANOVAs were conducted for each subexperiment.

Structural Factors In the first subexperiment (binding configurations (i) and (ii)), we found a large and highly significant main effect of *Ana* ($F_1(1,51) = 137.471$, $p < .0005$; $F_2(1,3) = 105.005$, $p = .002$). Anaphors (mean =

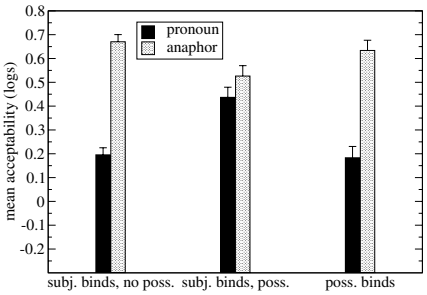


Figure 3: Structural effects on coreference judgments for binding in PNPs

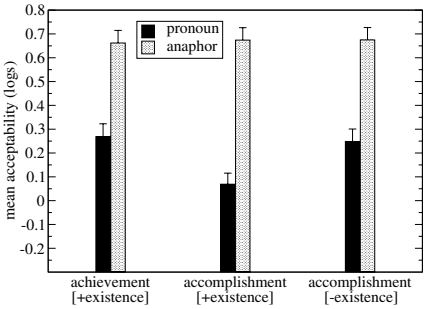


Figure 4: Effect of verb class on coreference judgments (subject binds, no possessor)

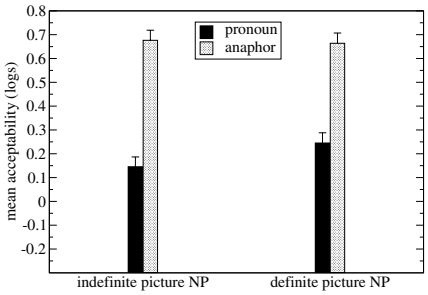


Figure 5: Effect of definiteness on coreference judgments (subject binds, no possessor)

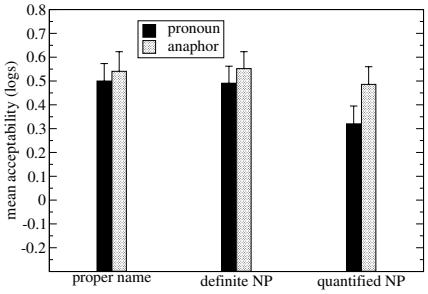


Figure 6: Effect of referentiality on coreference judgments (subject binds, possessor)

.6702) were more acceptable than pronouns (mean = .1954). In the second subexperiment (binding configurations (iii) and (iv)), *Ana* failed to reach significance: both anaphors (mean = .5262) and pronouns (mean = .4369) were equally acceptable. In the third subexperiment (binding configurations (v) and (vi)), again a main effect of *Ana* was present ($F_1(1,51) = 101.632$, $p < .0005$; $F_2(1,3) = 34.677$, $p = .010$). Anaphors (mean = .6338) were more acceptable than pronouns (mean = .1832). The coreference judgments for the six binding configurations (see Table 2) are graphed in Figure 3.

Pragmatic Factors The ANOVA for the first subexperiment also revealed a significant interaction of *Verb* and *Ana* ($F_1(2,102) = 11.275$, $p < .0005$; $F_2(2,6) = 6.193$, $p = .035$). This interaction is graphed in Figure 4, showing a decrease in the acceptability of pronouns for [existence] accomplishment verbs. An interaction of *Def* and *Ana* was also found, which however was significant by subjects only ($F_1(1,51) = 11.849$, $p = .001$; $F_2(1,3) = 2.168$, $p = .237$). Figure 5 shows that the acceptability of pronouns is increased for definite PNPs.

The ANOVA for the second subexperiment showed an interaction of *Ref* and *Ana*, significant by subjects only ($F_1(2,102) = 3.979$, $p = .049$; $F_2(2,6) = 2.745$, $p = .142$). This interaction is depicted in Figure 6, showing a decrease in the acceptability of pronouns if the antecedent is a quantified NP. No *Ref*/*Ana* interaction was present in the third subexperiment (see Figure 7).

Discussion

The theoretical predictions for the acceptability of the stimuli are listed in Table 2. Theory also predicts that anaphors are exempt from BT in configuration (i), and that structural factors should fail to have an influence on the acceptability of coreference for these structures. Contrary to this prediction, the present experiment revealed a significant influence of structural factors, although not in a way that any existing account predicts. Four major results were obtained.

In cases where the antecedent is in the subject and there is no possessor in the PNP (configurations (i) and (ii), see Table 2), structural and pragmatic binding theories alike predict that pronouns are fully acceptable and that pronouns and anaphors are equally acceptable. Our first and second major results are the falsification of both these predictions. Pronouns were significantly less acceptable than anaphors (see Figure 3). A comparison with standard cases of BT tested in Experiment 1 (see Figure 2) indicates that anaphors are fully acceptable in this configuration, while pronouns are of intermediate acceptability (but not fully unacceptable compared to, e.g., name-pronoun configurations with c-command).

Configurations (iii) and (iv), where the antecedent is in the subject, but there is a possessor, demonstrate our third major result. Here BT falsely predicts that anaphors

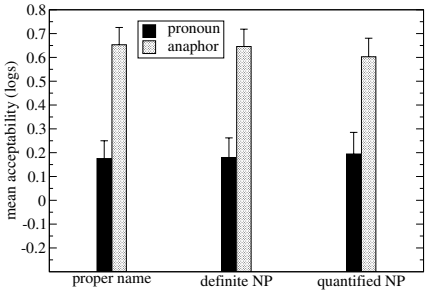


Figure 7: Effect of referentiality on coreference judgments (possessor binds)

are fully unacceptable. Note also that these anaphors are not exempt according to the pragmatic versions of BT, as there is a possessor. We found that pronouns and anaphors are both highly acceptable; no significant acceptability difference could be detected (see Figure 3). In other words, contrary to all that has been written in the syntactic literature, anaphors can be bound by the subject even in PNPs with possessors.

Our fourth result concerns the third structure we investigated (configurations (v) and (vi)), where the antecedent is the possessor in the PNP rather than the subject. We found the same behavior as in configurations (i) and (ii): the anaphors were fully acceptable in this configuration, while pronouns were significantly less acceptable, but not completely unacceptable compared to the configurations investigated in Experiment 1 (see Figure 2). The prediction for a PNP with a possessor is that a pronoun bound by the possessor is completely ungrammatical and that a pronoun bound by the subject is completely grammatical. This prediction was not supported by our results. We found that a pronoun bound by the possessor is as acceptable as a pronoun bound by a subject, but that both are only moderately acceptable.

We also investigated the influence of the pragmatic factors verb class, definiteness, and referentiality on coreference in PNPs. The underlying theoretical assumption is that coreference for exempt anaphors is governed by pragmatic factors, rather than by structural constraints. In binding configurations (i) and (ii), we found a significant effect of verb class: the acceptability of pronouns was reduced for [existence] accomplishment verbs (see Figure 4). This accords with intuitions reported in the literature (see Introduction, (4)). Furthermore, we found a significant effect of definiteness: pronouns are more acceptable with definite PNPs than with indefinite ones (see Figure 5). However, the verb class effect and the definiteness effect were weak and did not change the overall acceptability pattern, i.e., the preference for anaphors over pronouns.

In configurations (iii) and (iv), we found that the pragmatic factor referentiality has a significant effect on the acceptability of pronouns, which were less acceptable if the antecedent is a quantified NP, compared to cases where the antecedent is a name or a definite NP (see Figure 6). Again, this effect was weak and did not change the overall pattern, i.e., the fact that both pronouns and anaphors were highly acceptable in configurations (iii) and (iv). Finally, we failed to find any effect of referentiality in configurations (v) and (vi) (see Figure 7).

Conclusions

Experiment 1 was a control study that made a methodological contribution. The results showed that the experimental paradigm of magnitude estimation, previously only used for grammaticality judgment tasks, can be applied successfully to coreference judgments, which form the empirical basis of binding theory.

Building on this result, Experiment 2 used magnitude estimation to provide crucial data regarding binding in PNPs, which have been the subject of much research in the syntactic literature. The results provide an example of how experimentation can be used as a tool to settle debates in linguistic theory.

More specifically, Experiment 2 aimed to clarify the empirical status of exempt anaphors and provide data to distinguish between structural and pragmatic accounts of exempt anaphors. The results show that structural factors govern the binding possibilities in PNPs, while pragmatic factors play only a limited role. However, the structural factors identified differ from the ones standardly assumed. We found that (i) an anaphor can be bound from outside the PNP, even if there is a possessor in the PNP, (ii) anaphors and pronouns bound by the subject are equally acceptable when there is a possessor, (iii) pronouns are only moderately acceptable when there is no possessor, and (iv) pronouns bound by the possessor are also moderately acceptable.

Finding (i) is the most theoretically interesting one, and has recently been confirmed in an eye-tracking experiment (Runner, Sussman, & Tanenhaus, 2000). It falsifies a major prediction of all binding theories by showing that structural factors (subject/no subject, possessor/no possessor) fail to influence the binding of anaphors in PNPs. This means that the role of structural factors is even smaller than envisaged by proponents of pragmatic accounts. For pronouns, however, there is a structural effect, viz., they are more acceptably bound by the subject if there is a possessor NP.

In our view, the best way to understand this result is by making reference to the notion of predication (Pollard & Sag, 1994; Reinhart & Reuland, 1993). An anaphor must be bound by a dominating coargument of the predicate that selects for the anaphor, if there is one. For example, an anaphor that is in the object position of a matrix clause must be bound by the subject, because the subject position dominates the object position: both the subject and object are arguments of the same predicate, i.e., the predicate needs the subject and object to satisfy its syntactic and semantic requirements. But the possessor of a PNP is not an argument of the head, as the head does not require it (i.e., pictures do not *necessarily* have possessors). This observation correctly accounts for the full acceptability of anaphors in PNPs, with or without possessors, and the necessity for local binding when anaphors are in matrix argument positions (as in (1)).

We can also use the notion of predication to understand the pattern for [existence] accomplishment verbs, as in (4) *Hanna$_i$ took a picture of *her$_i$/herself$_i$*, without positing a covert possessor. It is possible that speaker-hearers treat expressions like *take a picture* as one predicate, in which case the anaphor or pronoun in such examples is actually a coargument of the subject and governed by Principle A or B, respectively. Runner (to appear) argues for just such an analysis of predicates like *take a picture* based on syntactic and semantic evidence.

References

Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, *72*(1), 32–68.

Chomsky, N. (1981). *Lectures on government and binding.* Dordrecht: Foris.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use.* New York: Praeger.

Chomsky, N., & Lasnik, H. (1995). The theory of principles and parameters. In *The minimalist program* (pp. 13–127). Cambridge, MA: MIT Press.

Corley, M., & Scheepers, C. (in press). Syntactic priming in English: Evidence from response latencies. *Psychonomic Bulletin*.

Cowart, W. (1997). *Experimental syntax: Applying objective methods to sentence judgments.* Thousand Oaks, CA: Sage Publications.

Gordon, P. C., & Hendrick, R. (1997). Intuitive knowledge of linguistic coreference. *Cognition*, *62*, 325–370.

Keller, F., & Alexopoulou, T. (2001). Phonology competes with syntax: Experimental evidence for the interaction of word order and accent placement in the realization of information structure. *Cognition*, *79*(3), 301–372.

Keller, F., Corley, M., Corley, S., Konieczny, L., & Todirascu, A. (1998). *WebExp: A Java toolbox for web-based psychological experiments* (Technical Report No. HCRC/TR-99). University of Edinburgh: HCRC.

Kuno, S. (1987). *Functional syntax: Anaphora, discourse and empathy.* Chicago: University of Chicago Press.

Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar.* Chicago: University of Chicago Press.

Reinhart, T., & Reuland, E. (1993). Reflexivity. *Linguistic Inquiry*, *24*, 657–720.

Runner, J. T. (to appear). When Minimalism isn't enough: An argument for argument structure. *Linguistic Inquiry*.

Runner, J. T., Sussman, R. S., & Tanenhaus, M. K. (2000). *Binding reflexives and pronouns in real-time processing.* Poster at the 13th Annual CUNY Conference on Human Sentence Processing, San Diego, CA.

Schütze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology.* Chicago: University of Chicago Press.

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects.* New York: John Wiley.

Training for Insight: The Case of the Nine-Dot Problem

Trina C. Kershaw (tkersh1@uic.edu) and Stellan Ohlsson (stellan@uic.edu)
Department of Psychology
University of Illinois at Chicago
1007 W. Harrison St., Chicago, IL 60607

Abstract

Three sources of difficulty for the nine-dot problem were hypothesized: 1) *turning on a non-dot point*, i.e., ending one line and beginning a new line in a space between dots; 2) *crossing lines*, i.e., drawing lines that intersect and cross; and 3) *picking up interior dots*, i.e., drawing lines that cross dots that are in the interior of the nine-dot and its variants. Training was designed to either facilitate or hinder participants in overcoming these difficulties. Participants were then tested on variants of the nine-dot problem. Results showed that participants in the facilitating training condition performed significantly better than the hindering or control group.

Constraints and Insights

Prior knowledge is the main resource that a problem solver can bring to bear on a problem. Prior knowledge produces unconscious biases that might influence perception and/or encoding of a problem. In general, prior knowledge can be helpful and productive when reasoning or solving a problem. However, when a problem solver faces a very unfamiliar or novel type of problem, there is no guarantee that prior knowledge will be relevant or helpful. The defining characteristic of so-called insight problems is that they activate seemingly relevant prior knowledge which is not, in fact, relevant or helpful (Ohlsson, 1984b, 1992; Wiley, 1998). To succeed, the problem solver must de-activate or relax the constraints imposed by the more or less automatically activated but unhelpful knowledge. To understand human performance on an insight problem, we should therefore try to identify the particular prior concepts, principles, skills or dispositions that constrain performance on that problem. Knoblich, Ohlsson, Haider, and Rhenius (1999) and Knoblich, Ohlsson, and Raney (1999) applied this perspective with considerable success to a class of match stick problems. In this paper, we apply it to the nine-dot problem and other connect-the-dots (CD) problems.

The nine-dot problem (Maier, 1930) requires that nine dots arranged in a square be connected by four straight lines drawn without lifting the pen from the paper and without retracing any lines (Figure 1). This task is ridiculously simple in the formal sense that there are only a few possible solutions to try, but ridiculously difficult in the psychological sense that the solution rate among college undergraduates who are given a few minutes to think about it is less than 5% (Lung & Dominowski, 1985; MacGregor, Ormerod, & Chronicle, 2001). The problem is surely of an unfamiliar type – when in everyday life do we ever draw lines to connect dots under certain constraints? – but what, exactly, are the sources of difficulty? Interestingly, seventy years of research (Maier, 1930) have not sufficed to answer this question.

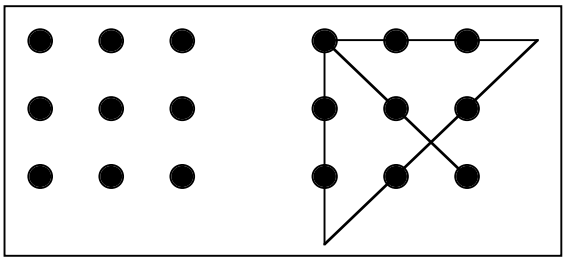


Figure 1: The nine-dot problem and its solution

The Gestalt psychologists introduced insight problems into cognitive psychology and explained their difficulty in terms of Gestalts, schemas that supposedly organize perceptual information (Ohlsson, 1984a). Consequently, they hypothesized that the nine-dot problem is difficult because people are so dominated by the perception of a square that the do not 'see' the possibility of extending lines outside the square formed by the dots (Scheerer, 1963). This hypothesis predicts that telling participants that they can draw lines outside the figure should facilitate the solution. Burnham and Davis (1969) and Weisberg and Alba (1981) tested this hypothesis, and found that the instruction only worked if combined with other hints that gave away part of the solution, e.g., telling the participants at which point to start or giving them the first line of the solution. A second prediction from the Gestalt hypothesis is that altering the shape of the problem and thus breaking up the square should also help. Both Burnham and Davis (1969) and Weisberg and Alba (1981) found facilitating effects of this manipulation. A third prediction from the Gestalt hypothesis is that giving people experience in extending lines outside the figure should help. Weisberg and Alba (1981) and Lung and Dominowski (1985) indeed found facilitating effects of such training.

Recently, MacGregor et al. (2001) and Chronicle, Ormerod, and MacGregor (in press) have proposed a

theory that attempts to predict quantitative differences in solution rates for different CD problems. Their explanation is based on four principles: (a) People always draw their next line so as to go through as many dots as possible. (b) People judge the value of a line as a function of how many dots it picks up in relation to how many dots are left and how many lines they have left to draw. (c) People look ahead 1, 2, 3 or at most 4 steps when deciding which line to draw next. (d) When lookahead indicates that every possible line from the current dot will end in a situation where the next line does not provide sufficient progress, they consider, with some probability, lines that go outside the figure formed by the dots.

This theory successfully predicts the differences in solution rates between different several different CD problems. It provides a more detailed description of why people get stuck than any previous theory – their lookahead is not deep enough to reveal that the solution path they are trying will dead end eventually – but the basic explanation for the difficulty is similar to that of previous theories: people consider lines within the shape formed by the dots before they consider lines that go outside the figure.

However, if this is true of variants of the nine-dot problems that do not form squares or any other 'good figure', then the Gestalt explanation for why people do not go outside the figure no longer holds. So what is the difficulty?

By analyzing pilot data and inspecting MacGregor et al.'s (2001) solution rates for the different nine-dot variants, we hypothesize that "hesitating to go outside the figure formed by the dots" is the wrong formulation of the constraint operating in this type of problem. Instead, we propose that *people are disposed to turn on a dot*, as opposed to turn on a point on the paper where there is no dot (a non-dot point). This constraint overlaps in meaning with the stay-within-the-figure constraint, so it explains the success of the training provided by Lung and Dominowski (1985). At the same time, this formulation is different enough to explain why telling people that they can go outside the figure does not help; they do not hesitate to extend lines outside the figure, but they do not want to turn on a non-dot point. As a secondary constraint, we hypothesize that people hesitate *to cross lines*, having a strong disposition towards thinking of the four lines they are supposed to draw as forming a closed outline. As a consequence, they do not see how *to pick up the dots in the interior of the figure*, an operation that requires crossing lines in many CD problems.

In the present study, we tested this hypothesis by both comparing problems that did and did not require turns on non-dot points and by attempting to facilitate the solution via training. As a novel methodological feature, we also tried to *hinder* the solution with training intended to *strengthen* the inappropriate constraints.

Method

Participants

Participants were 90 undergraduates (30 in each training group: facilitating, hindering, and no training) from UIC's Participant Pool. No demographic data were collected about the participants.

Materials

The training exercises were designed by the first author.

Facilitating Training The facilitating training was designed to eliminate the difficulties that participants were thought to face when solving the nine-dot problem. Twelve training exercises were designed, each with similar instructions to the nine-dot problem (Connect all of these dots using ___ straight lines without lifting your pen from the page and without retracing any lines). Each exercise required a different number of lines to connect the dots.

Six of the training exercises required participants to cross lines and pick up interior dots (Figure 2), and the other six could only be solved by turning on a non-dot point (Figure 3). Each training exercise was presented on its own page. The dots were filled circles that were .5 cm in diameter, and the centers of each dot were approximately 3.75 cm apart.

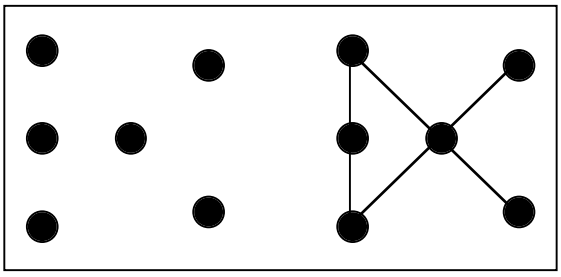


Figure 2: Facilitating Training Exercise and its Solution: Crossing Lines and Picking up Interior Dots

Hindering Training The hindering training consisted of 12 exercises that were solved by drawing lines that always turned on a dot and never crossed another line (Figure 4). As in the facilitating training, participants were instructed with similar directions to the nine-dot problem (Connect all of these dots using ___ straight lines without lifting your pen from the page and without retracing any lines). Again, each exercise required a different number of lines to connect the dots. The

hindering training was constructed just as the facilitating training, with the dots, or filled circles, being .5 cm in diameter and the centers of each dot being approximately 3.75 cm apart.

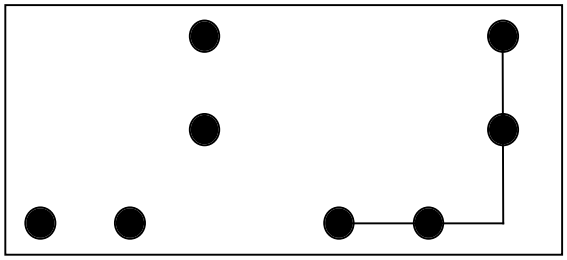


Figure 3: Facilitating Training Exercise and its Solution: Turning on a Non-Dot Point

Nine-Dot Variants The three insight and three non-insight versions of the nine-dot problem that were used had been designed by MacGregor et al. (2001). The insight problems required participants to turn on a non-dot point (Figure 5), while the non-insight problems were the insight problems with an added dot, which excused participants from having to turn on a non-dot point (Figure 6). Each problem was presented on its own page. The dots were filled circles that were 1 cm in diameter. The center of each dot was approximately 3.75 cm apart.

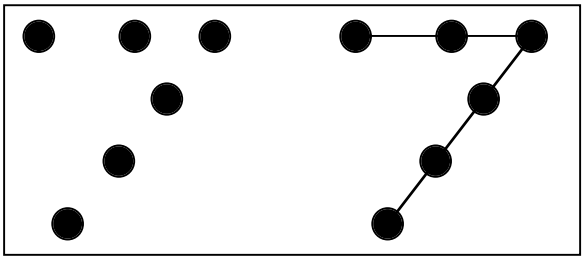


Figure 4: Hindering Training Exercise and its Solution

Procedure

Participants were seen in groups of 2-10. Each session lasted from 40 minutes to an hour. All test materials were contained in a booklet.

Training Phase Participants in both the facilitating and hindering training conditions were given the same directions for the training exercises. The instructions explained that they would be connecting dots using the number of lines specified on each page without lifting their pens from the page or retracing any lines. They were also told to start at the dot marked with a star for each group. The purpose of giving participants a set starting point was to make sure that there was a single solution for each training exercise.

Participants had one minute to work on each training exercise. Time was kept by the experimenter. Participants in the no training (control) group did not complete the training exercises and instead began with the problems.

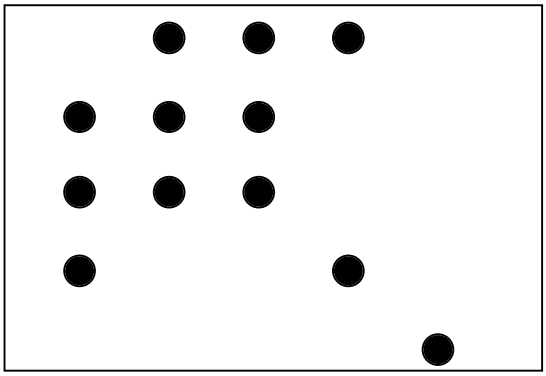


Figure 5: Nine-Dot Variant: Insight Version

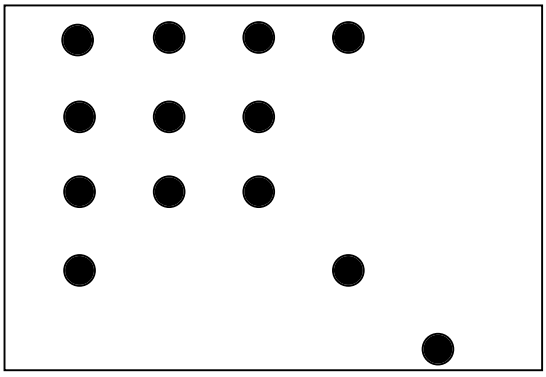


Figure 6: Nine-Dot Variant – Non-Insight Version

Problem-Solving Phase. After completing the training, participants began the problem-solving section of the booklet. Participants were instructed that they would have four minutes to connect all the dots in the figure using four straight lines without lifting their pens from the page and without retracing any lines. For each problem, there was a practice sheet followed by a second identical sheet with the problem on it. Participants were instructed to try the problem as many times as they wanted to on the practice sheet. When they thought they had come up with a solution, they were to record the time out of the four minutes that had passed, using a large clock at the front of the room.

The participants then were to turn the page and redraw their final solutions on the clean page.

The order of problems was the same for all participants. The first problem was an insight problem, the second was a non-insight problem, the third was an insight problem, and so on. Each insight problem was followed by its non-insight version. The first insight problem was of principal interest in comparing the effect of the training manipulations. The other problems were included to obtain baseline solving rates for various problem types, and to compare solving rates in the UIC participant population to the solving rates obtained by MacGregor et al. (2001).

Results

The results focus on the first insight problem (see Figure 5).

Group Analysis

A 3 x 2 chi-square analysis was conducted for the first insight problem. The independent variable was the type of training: facilitating, hindering, or no training. The dependent variable was whether or not a participant had solved the first insight problem. Nineteen participants (63%) in the facilitating group solved the first insight problem, eight (27%) in the hindering group solved, and 11 (37%) in the no training group solved. The chi-square was significant, $\chi^2(2, \underline{N} = 90) = 8.836$, $\underline{p} < .05$. Post-hoc comparisons showed significant differences between the facilitating group and the hindering group, $\chi^2(1, \underline{N} = 60) = 8.148$, $\underline{p} < .05$, and between the facilitating group and the no training group, $\chi^2(1, \underline{N} = 60) = 4.267$, $\underline{p} < .05$, but not between the hindering group and the no training group, $\chi^2(1, \underline{N} = 60) = .693$, $\underline{p} > .05$.

A between-groups analysis of variance (ANOVA) was conducted for the average amount of time it took participants in each group to solve the first insight problem. Participants in the facilitating group averaged 116 seconds to solve the first insight problem, the hindering group averaged 188 seconds, and the no training group averaged 185 seconds to solve. The ANOVA was significant, $\underline{F}(2,87) = 5.873$, $\underline{p} < .05$. Post-hoc Tukey tests revealed significant differences between the facilitating group and the hindering group ($\underline{p} < .05$) and the facilitating group and the no training group ($\underline{p} < .05$), but not between the hindering group and the no-training group ($\underline{p} > .05$).

Individual Differences Analysis

Although the facilitating group did better on the first insight problem than the other two groups, there was a large amount of variation in solving rate within the facilitating training group. Specifically, not all participants in the facilitating training group completed the training correctly. Participants in the facilitating group were split into two sub-groups based on whether or not they had completed the training correctly. In order to be classified as having completed the training correctly, a participant had to correctly complete over half (six) of the training exercises. If a participant did not correctly complete at least six of the training exercises, then he or she was put into the "did not complete training" group. In the "completed training" group, no participant got more than four training exercises incorrect. In the "did not complete training" group, one participant got six exercises wrong, and the others got seven or more exercises wrong.

There were a total of 19 participants in the "completed training" group, 17 (89%) of which solved the first insight problem, and 11 participants in the "did not complete training" group, two (18%) of which solved the first insight problem. A chi-square analysis comparing the performance of the two sub-groups on the first insight problem was significant, $\chi^2(1, \underline{N} = 30) = 15.248$, $\underline{p} < .05$.

Within each group, there were large differences in the amount of time needed to solve the first insight problem. Participants in the facilitating group who solved the first insight problem needed between 8 and 180 seconds to solve, with the majority of solvers requiring between eight and 109 seconds. The amount of time needed to solve ranged between 15 and 235 seconds in the hindering group, and between 20 and 195 seconds in the no training group.

Discussion

The results show that the problems that did not require turns on non-dot points were easier than those that did, and that the facilitating training improved performance on our CD problems, supporting the idea that the difficulties of the nine-dot problem and of CD problems generally might be some combination of a disposition towards turning on a dot and a disposition to think of the four lines they are supposed to draw as forming an outline and hence not crossing each other.

Contrary to expectation, the hindering training did not suppress the solution rate below that of a control group. There are several possible explanations. First, the solution rate was low enough so that attempting to suppress it further encountered a floor effect. Second, it is possible that the constraining dispositions were entrenched enough already so that attempting to entrench them yet further with a brief intervention did not succeed.

An interesting finding was that for the facilitating group, the degree to which participants completed the training determined their success in solving the first

insight problem. It is likely that only the participants in the facilitating group who fully completed the training were able to successfully transfer what they had learned during training to solving the insight problems. This finding shows that despite common difficulties for participants in solving CD problems, there exists individual variation in the degree to which participants can be guided to overcome these difficulties.

Where would people acquire the two central dispositions to want to turn on a dot and to draw outlines? The first disposition might stem from yet another Gestalt concept, the difference between figure and ground. The dots on the paper is the figure and the paper is the background and hence not part of what they are working on. The disposition to draw outlines might be grounded in how people draw when they try to make representational drawings; they trace the outline of the object they are trying to represent. Even if plausible sources can be identified for these two constraints, it remains to prove that they are operating in the nine-dot problem itself as well as in the altered versions we used in this study.

Another reason that people find the nine-dot and related CD problems difficult is that their prior experience in solving CD problems is based in children's connect-the-dot puzzles. This experience is irrelevant to the knowledge that is needed to solve the nine-dot problem. Prior knowledge creates unconscious biases that are not always helpful (cf. Ohlsson, 1984b, 1992; Wiley, 1998). The presentation of a problem can interact with prior knowledge, thus resulting in an incorrect and unhelpful encoding of the problem.

What is of most interest in studies on CD problems is to comprehend how people can get stuck on such trivial problems. To understand how the human mind works, we must understand unhelpful interactions between problems and prior knowledge, the impasses that result, and how people overcome those impasses by relaxing the inappropriate constraints. Insight problems are tools with which to study these processes.

References

Burnham, C.A., & Davis, K.G. (1969). The nine-dot problem: Beyond perceptual organization. Psychonomic Science, 17(6), 321-323.

Chronicle, E.P., Ormerod, T.C., & MacGregor, J.N. (in press). When insight just won't come: The failure of visual cues in the nine-dot problem. Quarterly Journal of Experimental Psychology.

Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. Journal of Experimental Psychology: Learning, Memory, and Cognition, 25, 1534-1555.

Knoblich, G., Ohlsson, S., & Raney, G. (1999). Resolving impasses in problem solving: An eye movement study. In M. Hahn and S. Stoness (Eds.), Proceedings of the Twenty-First Annual Meeting of the Cognitive Science Society (pp. 276-281). Mahwah, NJ: Erlbaum.

Lung, C.T., & Dominowski, R.L. (1985). Effects of strategy instructions and practice on nine-dot problem solving. Journal of Experimental Psychology: Learning, Memory, and Cognition, 11(4), 804-811.

MacGregor, J.N., Ormerod, T.C., & Chronicle, E.P. (2001). Information-processing and insight: A process model of performance on the nine-dot and related problems. Journal of Experimental Psychology: Learning, Memory, and Cognition,27(1), 176-201.

Maier, N.R.F. (1930). Reasoning in humans: I. On direction. Journal of Comparative Psychology, 10, 115-143.

Ohlsson, S. (1984a). Restructuring revisited I: Summary and critique of the Gestalt theory of problem solving. Scandinavian Journal of Psychology, 25, 65-78.

Ohlsson, S. (1984b). Restructuring revisited II: An information processing theory of restructuring and insight. Scandinavian Journal of Psychology, 25, 117-129.

Ohlsson, S. (1992). Information processing explanations of insight and related phenomena. In M. Keane and K. Gilhooly (Eds.), Advances in the Psychology of Thinking (Vol.1, pp. 1-44). London: Harvester-Wheatsheaf.

Scheerer, M. (1963) Problem solving. Scientific American, 208(4), 118-128.

Weisberg, R.W., & Alba, J.W. (1981). An examination of the alleged role of "fixation" in the solution of several "insight" problems. Journal of Experimental Psychology: General, 110(2), 169-192.

Wiley, J. (1998). Expertise as mental set: The effects of domain knowledge on creative problem solving. Memory & Cognition, 26(4), 16-730.

Theory-Based Reasoning in Clinical Psychologists

Nancy S. Kim (nancy.s.kim@yale.edu)
Department of Psychology, Yale University; 2 Hillhouse Avenue
New Haven, CT 06520 USA

Woo-kyoung Ahn (woo-kyoung.ahn@vanderbilt.edu)
Department of Psychology, Vanderbilt University; 534 Wilson Hall
Nashville, TN 37240 USA

Abstract

Progress in science is marked by the formation of theories that explain a body of observations. Contrary to this guiding philosophy, clinical psychologists have prescribed to an atheoretical system of diagnostic reasoning since 1980. We report two studies demonstrating that clinicians have not internalized atheoretical reasoning despite the system's widespread acceptance. The results show that clinicians' own theories about disorders significantly affect their diagnoses of hypothetical patients and memory for symptoms. Clinicians are cognitively driven to form and apply theories to observations despite decades of deliberate training, practice, and pressure to reason atheoretically.

Introduction

Philosophers of science have argued for decades that scientific progress is delineated not merely by the amassing of observations, but rather by the formation of broad principles that organize and explain these observations in a cohesive manner (Hempel, 1965). Evolutionary theory, for example, is regarded as a revolutionary advance in organismic biology in large part because it provides a deeper structure for a scientific taxonomy of living things, moving away from classification based on superficial features. Data are measured and considered within the larger structure of this overarching theory. Similarly, the human mind constantly seeks out underlying rules and principles that make sense of incoming data concerning the surrounding world. Adults (Murphy & Medin, 1985), children (Gelman, 2000; Keil, 1989), and even infants (Marcus, Vijayan, Rao, & Vishton, 1999) spontaneously extract and apply underlying organizing principles and abstract rules that go beyond surface features. In this way, the human mind forms categories and concepts based on its theories about the surrounding world (Carey, 1985).

Atheoretical versus theory-based reasoning

In contrast, the current *Diagnostic and Statistical Manual of Mental Disorders* (APA, 1994), prescribes an atheoretical approach to diagnosing mental disorders (Follette & Houts, 1996). Most mental disorders lack a single universally acknowledged pathogenesis, which in the past led to unreliability between clinicians in diagnosis. The *DSM-IV*'s (APA, 1994) widely acclaimed solution is to eliminate theory use altogether when making a diagnosis, incorporating instead checklists of symptoms compiled by a panel of experts. In doing so, it represents each disorder as a list of unrelated symptoms, ignoring the causal relations between symptoms that are a fundamental aspect of theory representations (Carey, 1985). For most disorders, the *DSM-IV* (APA, 1994) states that a subset of the list is sufficient for a diagnosis regardless of which combination of symptoms appears, thereby assuming that all symptoms in the list are equally central to the disorder. For example, any 2 of the following 5 symptoms warrant a diagnosis of schizophrenia, according to the *DSM-IV*: hallucinations, delusions, disorganized speech, grossly disorganized or catatonic behavior, and negative symptoms. Since eliminating any overt mention of an underlying theory for the taxonomy two decades ago (APA, 1980), the *DSM* system has become widely accepted in the U.S., forming the core of research, clinical assessment, diagnosis, and treatment in psychopathology. Research funding, journal titles, and health care reimbursements are all organized by, and dependent on, use of the categories defined by the *DSM-IV* (APA, 1994).

Has the *DSM* system succeeded in internalizing atheoretical reasoning in clinicians? An atheoretical approach would suggest that experienced clinicians, after years of emphasizing use of the *DSM* system, will come to embody its prescription of atheoretical reasoning (APA, 1994). In contrast, the theory-based approach would suggest that clinicians, despite such emphasis on the elimination of theory, are still influenced by their own idiosyncratic theories about disorders when reasoning about them (Medin, 1989).

The two approaches were differentiated by testing for the presence or absence of the causal status effect, a specific mechanism by which theory-based reasoning occurs (Ahn, 1998; Ahn, Kim, Lassaline, & Dennis,

2000). The causal status effect is said to occur when category features causally central to an individual's theory of that category are treated as more important in categorization than less causally central features. For instance, if symptom A causes symptom B in a clinician's theory, then A is more causally central than B, and A is thereby predicted to have greater diagnostic importance than B. To derive the causal centralities of individual symptoms imbedded in a complex theory, the following formula[1] can be used:

$$c_{i,t+1} = \Sigma_j \, d_{ij} \, c_{j,t} \quad (1)$$

where d_{ij} is a positive number that represents how strongly symptom j depends on symptom i, and $c_{j,t}$ is the conceptual centrality of feature j, at time t (Sloman, Love, & Ahn, 1998). This model states that the centrality of feature i is determined at each time step by summing across the centrality of every other feature multiplied by that feature's degree of dependence upon feature i. Thus, in the current studies we operationalized the theory-based view as a systematic effect of relational structures on conceptual representation and use.[2]

A theory-based view would further predict that any features relationally connected to other features would be treated as more important than isolated features in reasoning (Gentner, 1983). That is, if symptom A causes symptom B, but symptom C is isolated (it does not and is not caused by any other symptoms in a clinician's theory), C would be the least central symptom of the three.

Study 1

We measured expert and trainee clinical psychologists' causal theories in the first session. Then we examined whether the causal centralities of symptoms in their theories predict how important these symptoms are in diagnosis (in the first and second sessions), and how well they are remembered (in the second session).

Participants

Participants were 11 experienced clinical psychologists, and 10 clinical psychology graduate students. The experienced clinical psychologists had been in practice for a minimum of 15 years (ranging from 15 - 52 years with a median of 28 years). Ten were licensed psychologists with Ph.D.'s, and 1 was a board-certified psychiatrist with an M.D.

Session 1: Measurement of causal theories and conceptual centrality

We measured participants' individual causal theories for each of 5 disorders that were judged to be highly familiar by undergraduate students. The 5 disorders were Anorexia Nervosa, Antisocial Personality Disorder, Major Depressive Episode, Specific Phobia, and Schizophrenia.

In an initial disorder defining task, participants viewed a list of symptoms for each disorder. Symptoms included both the *DSM-IV* (APA, 1994) diagnostic criteria and the non-criterial, characteristic symptoms from the manual's disorder description. Participants were asked to define each disorder for themselves by adding new symptoms, crossing out symptoms, combining two or more symptoms, and / or dividing a single symptom into two or more symptoms.[3] All subsequent tasks in both sessions incorporated these individually tailored lists.

Participants' causal theories were then measured for each disorder. Participants received slips of paper, each bearing the name of a symptom. They were first asked to arrange the symptoms around the corresponding disorder name. Next, participants drew arrows between symptoms to indicate causal relations as they thought was appropriate. Finally, they rated the strength of each causal relation on a scale of 1-5 (1=very weak; 5=very strong). From these causal drawings, we determined the causal centrality of each symptom using Equation (1). Isolated features were always assigned the lowest causal centrality.

During this session, we also measured the conceptual centrality of each symptom to the disorder. Clinicians were asked, "how easily can you imagine a person with [disorder X] who does not have the symptom of [Y]?" for each symptom on a scale of 0-100 (0=very difficult to imagine; 100=very easy to imagine). The order of the two tasks, conceptual centrality and causal theory measurement, was counterbalanced between participants. The results demonstrated that conceptual

[1] Although other formulas are also consistent with the causal status effect, this formula showed the best fit in analyses of lay people's conceptual representations of common objects (e.g., apples and guitars; Sloman et al., 1998). Moreover, all of the analyses on causal centrality reported below are based on rank orders of causal centrality derived from this formula, and different formulas do not produce radically different rank orders.

[2] We do not intend to claim here that theory-based categorization is limited to the effect of relational structures. Categorization may also be affected by the content of relations, an issue that was not the focus of the current studies.

[3] For instance, a participant might choose to divide the single symptom "disturbed experience of body shape or denial of the problem" into separate symptoms ("disturbed experience of body shape;" "denial of the problem").

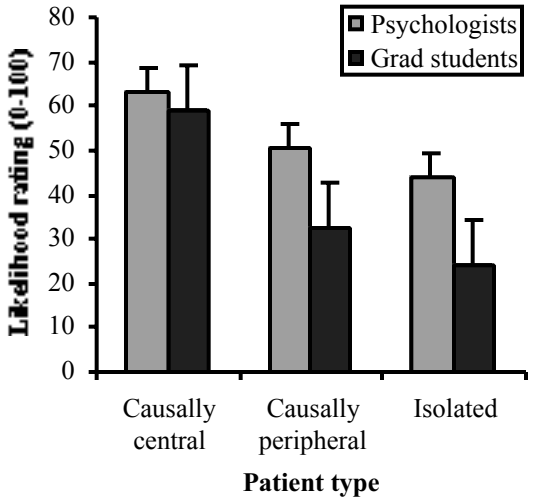


Figure 1. Clinical psychologists' and clinical graduate students' likelihood ratings of mental disorder category membership for hypothetical patients in Study 1. Error bars indicate standard errors.

centrality was positively correlated with causal centrality (as determined by Equation [1]) for 18 out of 20 participants,[4] such that the more causal a symptom was within each participant's theories, the more central the symptom was to that participant's concept of the disorder. The median overall correlation coefficients for clinical psychologists and graduate students were .41 (range: -.12 to .50) and .27 (range: .11 to .62), respectively.

Session 2: Hypothetical patient diagnosis and free recall of symptoms

Participants were brought back to the lab about 10 - 14 days after the first session. For the second session, we created a unique set of hypothetical patients for each participant based on that participant's own theories as reported in the causal centrality task. Two to three hypothetical patients were constructed for each of the 5 mental disorder categories. Each patient was composed of three symptoms that were either causally central or causally peripheral as determined by Equation (1), or isolated[5]. For example, causally central patients consisted of three symptoms causally central to the participant's theory of the disorder (i.e., they were thought to cause more symptoms more strongly than other symptoms did). Participants were told that these patients did not exhibit any other symptoms. The number of *DSM-IV* (APA, 1994) diagnostic criteria was equated over the types of patients. Therefore, according to strict *DSM-IV* (APA, 1994) criteria, each of the three different types of hypothetical patients should be equally likely to have the disorder in question.

We asked participants to rate the likelihood that each of these hypothetical patients actually had the associated disorder. Specifically, for each hypothetical patient, participants answered the question, "what is the likelihood, in your opinion, that a patient with the following characteristics has [disorder X]?" on a scale of 0-100 (0=very unlikely; 100=very likely).

Although the number of criterial symptoms according to the *DSM-IV* (APA, 1994) was equated across the three types of patients, participants judged patients with causally central symptoms (mean of 61.0) as nearly 20% more likely to have the disorder than patients with causally peripheral symptoms (mean of 42.0; $t = 4.5$, $p < .001$). Patients with isolated symptoms were judged as least likely to have the disorder (mean of 34.5; $t = 3.3$, $p = .003$). Figure 1 shows the results broken down by expert and trainee participant groups. There was neither a significant main effect of expertise nor any significant interaction involving expertise.

Approximately one hour after they completed the hypothetical patient diagnosis task, participants were asked to recall the symptoms of those hypothetical patients. Participants recalled significantly more causally central (67%) than causally peripheral (51%; $t = 2.9$, $p = .009$) or isolated (44%; $t = 2.7$, $p < .02$) symptoms.

Study 2

Study 2 expanded the generality of these findings using modified procedures. There were two principal changes. First, the causal centrality task was modified to measure participants' theories about all kinds of symptom-symptom relations, not restricting the measure to causal relations only. Second, another aspect of memory for symptoms was examined by using a recognition task instead of the recall task. These changes will be described in detail in the following sections.

Participants

Participants were another group of 14 experienced clinicians and 6 clinical psychology interns. The expert clinicians had been in practice for a minimum of 15 years (ranging from 17-43 years with a median of 26 years). All 14 expert clinicians were licensed psychologists; 13 had Ph.D.'s and 1 had an Ed.D.

[4] The data of one graduate student participant could not be included in this particular analysis because there was no variance among that participant's conceptual centrality responses within each disorder.

[5] Some participants left no symptoms isolated in their theory of a disorder. In these cases, a patient description composed of isolated symptoms could not be created.

Stimulus materials

The same 5 disorders in Study 1 were also utilized in this study. Unlike in Study 1, however, participants were provided with a list of "standard" symptoms. These symptoms were defined by the participants in Study 1. Namely, a symptom was dropped from the list of symptoms for Study 2 if it was dropped by over 50% of the experts and over 50% of the trainee participants in Study 1. Using these standard lists of symptoms allowed us to make direct comparisons between participants' theories, especially between those of experts and trainees.

Session 1: Measurement of relational theories and conceptual centrality

We measured participants' individual theories for each disorder using the same procedure as before, except that this time we asked participants to draw any kind of relations between symptoms they saw fit, not limiting the measure to causal relations. Participants rated the strength of each relation on a scale of 1-3 (1= weak; 2=moderate; 3=strong). They were asked to consider using, but not to limit themselves to, the following relations: "is a subset of," "is an example of," "precedes," "co-occurs with," "is a precondition for," "causes," "jointly cause," "affects," "determines the extent of," "increases," "decreases," "is a catalyst for," "is used as a defense against," "is a cure for." The relational centrality of each symptom was then determined using Equation (1). For instance, in applying the formula, "A is a precondition for B" and "A precedes B" are treated as "B depends on A." We used these relational centralities to further investigate clinicians' and clinical trainees' use of theories in reasoning with three modified measures of diagnostic importance and memory. During the first session, we also measured the diagnostic importance of each symptom to the disorder. Participants were asked, "how important is the symptom of [Y] in diagnosing a person with [disorder X]?" on a scale of 0-100 (0=very unimportant; 100=very important). As in Study 1, the order of the two tasks was counterbalanced between participants. The results demonstrated that diagnostic importance was positively correlated with relational centrality (as measured by Equation [1]) for 18 out of 20 participants (average r = .77, clinical psychologists; r = .66, interns). That is, the more symptoms depended on a symptom, the more central that symptom was to participants' concepts of the disorder. The median overall correlation coefficients for psychologists and graduate students were .25 (range: .02 to .61) and .44 (range: -.94 to .68), respectively.

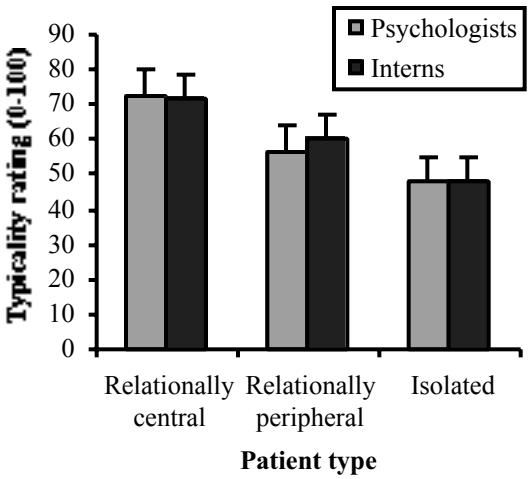


Figure 2. Clinical psychologists' and clinical psychology interns' typicality ratings for hypothetical patients in Study 2. Error bars indicate standard errors.

Session 2: Hypothetical patient typicality and recognition of symptoms

As in Study 1, we constructed 2 to 3 hypothetical patients for each of 5 mental disorder categories. Each patient consisted of a set of three symptoms that were either relationally central, relationally peripheral, or isolated, for each disorder and each participant. As in Study 1, a patient composed of isolated features was not included if a participant did not leave any symptoms isolated in their theory. Again, the number of criterial symptoms was equated between patients so that diagnoses based strictly on the *DSM-IV* (APA, 1994) would not differentiate them. Participants were asked to assess how typical hypothetical patients were of the disorder (following Cantor, Smith, French, & Mezzich, 1980). Specifically, participants answered the standard typicality rating question, "how well, in your opinion, does a patient with the following characteristics fit in the diagnostic category of [X]?" on a scale of 0-100 (0=very poorly; 100=very well) for each patient. Patients with relationally central symptoms (mean of 72.3) were judged as more typical of the disorder than patients with relationally peripheral symptoms (mean of 57.8; $t = 4.3$, $p < .001$), which in turn were judged as more typical than patients with isolated symptoms (mean of 47.9; $t = 2.7$, $p < .02$). Figure 2 shows the results broken down by experts and trainees. Neither a significant main effect of expertise nor an interaction effect involving expertise was found.

Following an approximately one-hour delay, participants received a standard recognition task (following Roediger & McDermott, 1995), in which they were asked to classify symptoms on a list as "old" or new" based on whether they had seen them earlier in the hypothetical patients task. The list included 30

relationally central and 30 relationally peripheral symptoms. Half of the symptoms in each group were old and half new. Consistent with previous findings showing an effect of schema on false recognition (Bower, Black, & Turner, 1979), participants were much more likely to falsely recognize new, relationally central symptoms as symptoms that they had seen before (23.3%) than new, relationally peripheral symptoms (13.2%; t = 3.0, p = .008). Participants showed greater sensitivity to relationally peripheral symptoms (d' = 2.62) than to relationally central symptoms (d' = 2.04; F[1, 18] = 2.43; p = .01). Thus, participants were less able to distinguish between presented and non-presented relationally central symptoms. False memory has generally been thought to be an issue for patients with psychological disorders or problems (Loftus & Ketcham, 1994). Interestingly, we found that therapists are biased to falsely remember having seen symptoms in their patients that are central to their personal theories about the disorder.

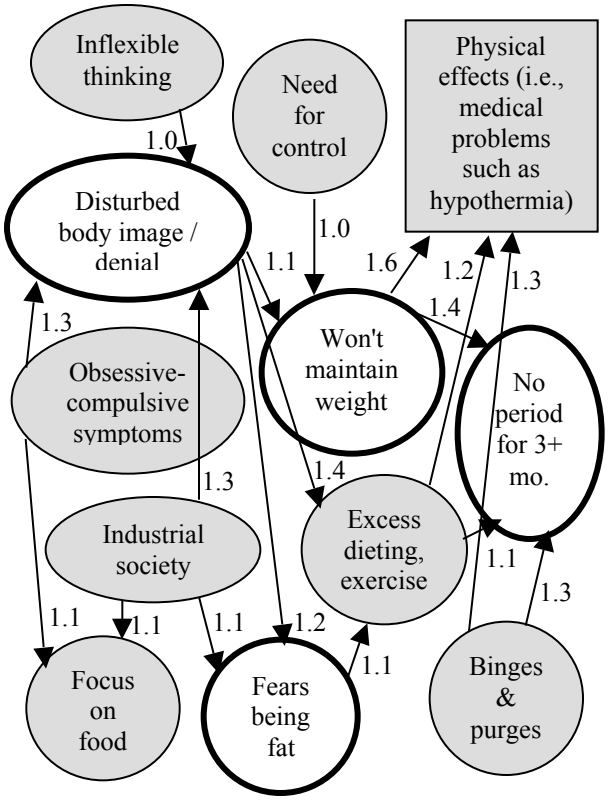


Figure 3. Participants' average relational theory of Anorexia Nervosa in Study 2. (Note: An arrow from A to B indicates that B depends on A [or A affects B]. The symptom descriptions in the figure were truncated and sometimes combined to keep the figure readable. Only dependency relations that received mean strength ratings greater than 1 on a 3-point scale are shown in the figure. Symptoms in white circles with boldface borders are *DSM-IV* (APA, 1994) diagnostic criteria. The *DSM-IV* implies that these are not weighted differently.)

Consensus on theories

Unlike in Study 1, all participants received the same set of symptoms, allowing a direct comparison of their theories. Participants' theories, as measured by relational centrality rank orders, were highly consistent with each other (Kendall's coefficients of concordance ranging from .34 to .59 across the 5 disorders, all p's < .0005). Because of this, we were able to construct an average dependency structure for each disorder, such as the one shown in Figure 3 for Anorexia Nervosa.

These average theories of experts generally agreed with those of lay people. For instance, the mean rank orders of diagnostic criteria symptoms for major depression obtained in an earlier study from undergraduate students (Kim & Ahn, 2001) were highly correlated with those obtained from clinicians and clinical trainees in Study 2 (r = .93, p < .001). We also developed hypothetical patient descriptions based on clinicians' and clinical trainees' averaged theories of the 5 disorders and gave them to 23 undergraduates. Relationally central patients (mean of 75.3) were judged to be more typical of a disorder than relationally peripheral patients (mean of 29.3; t = 12.44, p < .001).

Discussion

In sum, we found that symptoms playing a central role in clinicians' theories were considered to be more important in diagnosis, were more likely to be recalled later, and were more likely to be falsely recognized as having been present in patient descriptions. Despite the fact that the modern *DSM* system, which has become engrained into the conscious practice of clinical psychology, avoids the use of theory, clinicians prefer to base their reasoning on their own organizing theories. This was shown at a much more specific level of analysis than previous work documenting top-down effects of theory on reasoning (i.e., Chapman & Chapman, 1967; Wisniewski & Medin, 1994). Furthermore, in all six measures, there were no differences between levels of expertise, suggesting that years of training and long-term use of the modern *DSM* system do not diminish the effect of theory on reasoning.

Implications

When making formal *DSM* diagnoses using checklists, it is possible that clinicians may not be as strongly affected by their theories. However, the effect of theory-based conceptual representations found in the current studies may still pervade critical aspects of clinical work. As shown in our study, clinicians are

better at recalling symptoms central to their theories, and may be biased to falsely remember theory-central symptoms of patients they have already seen. These tendencies may influence clinicians' informal initial diagnoses, which may in turn markedly affect how clinicians subsequently perceive and interact with their patients. For instance, symptoms of mental disorders are often ambiguous, and clinicians may focus their attention on detecting symptoms central to their theories.

We also note that theory-based reasoning in itself is not a reasoning fallacy, provided that clinicians' theories are valid (Dawes, Faust, & Meehl, 1989). However, in the case of less well-known disorders such as personality disorders, experts may have more idiosyncratic theories. This, if true, may account in part for the notoriously low reliability between clinicians in diagnosing the personality disorders. We are currently conducting a study to examine this issue. In cases such as these, reliance on invalid theories may perhaps constitute a fallacy in clinical judgment.

In general, however, categorization based on valid theories conforms to the higher levels of taxonomy that scientists should strive for (Hempel, 1965). Indeed, symptoms that explain and cause other symptoms may be the most important ones to attend to and remember, because they may be the more useful predictors for prognosis and treatment. In the current study, we found clinicians' theories to be in general agreement with each other's and with lay people's theories, at least in disorders that are also familiar to lay people. This suggests that experts' theories of these socio-culturally familiar disorders are not highly idiosyncratic, but rather seem to be based on commonsense notions, and may therefore be worthy of careful consideration in revising the *DSM*.

Acknowledgments

We thank Marvin Chun, Frank Keil, Donna Lutz, Laura Novick, Peter Salovey, and Andrew Tomarken for helpful comments, and Jessecae Marsh and Judy Choi for help in running participants. This research was supported in part by a National Science Foundation Graduate Research Fellowship to Nancy S. Kim and a National Institute of Mental Health Grant (RO1 MH57737) to Woo-kyoung Ahn.

References

Ahn, W. (1998). Why are different features central for natural kinds and artifacts? *Cognition, 69*, 135-178.

Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology, 41*, 361-416.

American Psychiatric Association (1980). *Diagnostic and Statistical Manual of Mental Disorders*, 3rd Edition. Washington, DC: Author.

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders*, 4th Edition. Washington, DC: Author.

Bower, G. H., Black, J. B., Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology, 11*, 177-220.

Cantor, N., Smith, E. E., French, R., & Mezzich, J. (1980). Psychiatric diagnosis as prototype categorization. *Journal of Abnormal Psychology, 89*, 181-193.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: Plenum.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology, 72*, 193-204.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668-1674.

Follette, W. C., & Houts, A. C. (1996). Models of scientific progress and the role of theory in taxonomy development: A case study of the *DSM*. *Journal of Consulting and Clinical Psychology, 64*, 1120-1132.

Gelman, S. A. (2000). In H. W. Reese (Ed.), *Advances in child development and behavior*. San Diego, CA: Academic Press.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155-170.

Hempel, C. G. (1965). *Aspects of scientific explanation*. New York: Free Press.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Kim, N. S., & Ahn, W. (in press). The influence of naive causal theories on lay concepts of mental illness. *American Journal of Psychology*.

Loftus, E., & Ketcham, K. (1994). *The Myth of Repressed Memory*. New York: St. Martin's Press.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by 7-month-old infants. *Science, 283*, 77-80.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist, 44*, 1469-1481.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289-316.

Sloman, S. A., Love, B. C., & Ahn, W. (1998). Feature centrality and conceptual coherence. *Cognitive Science, 22*, 189-228.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*, 803-814.

Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science, 18*, 221-281.

Effect of Exemplar Typicality on Naming Deficits in Aphasia

Swathi Kiran (s-kiran@northwestern.edu),
Department of Communication Sciences and Disorders, Northwestern University,
Evanston, IL 60208 USA

Cynthia K. Thompson (ckthom@northwestern.edu)
Department of Communication Sciences and Disorders & Neurology, Northwestern University,
Evanston, IL 60208 USA

Douglas L. Medin (medin@northwestern.edu)
Department of Psychology, Northwestern University
Evanston, IL 60208 USA

Abstract

The effect of typicality of category exemplars on naming performance was investigated using a single subject experimental design across participants and behaviors in four patients with fluent aphasia. Participants received a semantic feature treatment to improve naming of either typical or atypical examples, while generalization was tested on the untrained examples of the category. The order of typicality and category trained was counterbalanced across the participants. Results indicated that patients trained on naming of atypical examples demonstrated generalization to naming of intermediate and typical examples. Patients trained on typical examples demonstrated no generalization to naming of intermediate or atypical examples. Implications for models of typicality and rehabilitation of aphasia are discussed.

Introduction

Aphasia is a language disorder that results from damage (such as stroke or head trauma) usually to the left hemisphere of the brain. Naming difficulty is the most common form of language deficit noted in individuals with aphasia. One widely accepted model of naming (Dell, 1986; Stemberger, 1985) suggests that activation of a word during naming involves two closely interacting levels, activation of the semantic representation as well as activation of the phonological form of the target word. Naming deficits can therefore arise from difficulty in activation at either of the two levels. Patients with naming deficits arising from an impairment in activating semantic representations often present with impairments in accessing appropriate semantic fields within categories (Goodglass & Baker, 1976).

Numerous studies on normal individuals have found typical examples of a category to be accessed faster and more accurately than atypical examples, an effect labeled the typicality effect (Rips, Shoben, & Smith, 1973; Rosch, 1975). Evidence for the typicality effect exists through typicality ratings (Rosch, 1975), response times on category verification tasks (Larochelle & Pineu, 1994), and category production frequency (Rosch, 1975). Little evidence, however, exists regarding representation of typicality in individuals with aphasia, although some investigations have noted deficits (Grober et al., 1980; Grossman, 1981). The interpretation of these deficits with reference to theoretical models of typicality however has not been addressed.

In a connectionist account of relearning in neural networks, it was found that a lesioned computer network retrained on atypical examples resulted in improvements on typical items as well (Plaut, 1996). Training typical items, however, only improved the performance of those items while performance of atypical words deteriorated during treatment. While Plaut's findings have not yet been tested in individuals with aphasia, the prospect of such generalization effects is especially significant for treatment of naming deficits, since most naming treatments have found little generalization to untrained items (McNeil et al., 1997; Pring et al., 1993).

The present experiment aimed to investigate the effects of exemplar typicality on naming performance in individuals with aphasia. Specifically, the purpose of the experiment was to train naming of a set of typical or atypical examples of a superordinate category, and examine generalization to untrained examples of the category. The present experiment was motivated by prototypical/family resemblance models of typicality (Hampton, 1979; Rosch & Mervis, 1975). According to these models, on a multidimensional scaling of a category (e.g., bird) based on similarity of items, typical examples (e.g., robin, sparrow) are found to have more features similar amongst them and with the category prototype, and therefore are represented closer to the center of the semantic space. In contrast,

atypical items (e.g., penquin, ostrich) have fewer features that are similar amongst them and the prototype, and are at the periphery of this semantic space. We hypothesized that training aphasic individuals to produce atypical examples from a category would result in generalization to more typical examples of the category. If indeed atypical examples are at the periphery of the category, then strengthening access to these examples by emphasizing the variation of semantic features across the category would strengthen the overall semantic category. Conversely, typical examples were hypothesized to represent little or no variation within the category. Therefore, training typical examples was predicted to improve only items at the center of the category, with no improvements expected for atypical examples.

Methods

Participants

Four individuals, ranging in age from 63-75 years, and presenting with aphasia resulting from a cebrebrovascular accident to the left hemisphere were selected for the experiment. All four patients presented with fluent aphasia, characterized by fluent circumlocutory speech, mild auditory comprehension deficits and severe naming difficulties. Based on standardized language testing, the locus of naming deficit was attributable to impairments in accessing the semantic representation of the target, and/or in accessing its phonological form.

Stimuli

Norms for typicality of category exemplars were developed prior to initiation of the experiment. One group of 20 normal young and elderly subjects constructed as many examples as possible for ten categories, while another group of 20 normal young and elderly subjects rated the typicality of these examples on a 7-point scale. Examples for each category were then divided into three groups, typical, intermediate and atypical, based on their average z scores. Based on several selection criteria, which included frequency, distinctiveness, number of syllables, unambiguity regarding category membership, two categories (birds and vegetables) with 24 examples each were selected for treatment. Each set of 24 items included a subset of eight typical and eight atypical items. The remaining eight in each set were determined to be intermediate in terms of typicality. For each of the selected examples, corresponding color photos printed on 4 x 6 inch cards were selected. In addition to the experimental photos, stimuli from three different superordinate categories (fruit, animal and musical instrument) were selected to serve as distracters for treatment.

Once the two categories and their 24 examples were selected, semantic features for each category were developed. For each category, a minimum of 20 features belonging to the category that were either physical, functional, characteristic or contextual attributes were selected. Additionally, a minimum of 20 distracter features to be used during the yes/no question tasks (see treatment), using the same four attribute types not belonging to the target category were developed. At least 10 features that were applicable to all examples in the category were selected (e.g., bird: has a beak, lays eggs), while obscure features (e.g., asian food for vegetable), and features that were salient only for a single example (e.g., hoots for owl, drills holes for wood pecker) were eliminated. Generally, features that were applicable to two or more items in the category were selected. Distracter features belonging to the categories sport, transportation, animal, insect, flower and weapon were selected using the same criteria as the target category features.

Design

A single subject experimental design with multiple baselines across behaviors and participants (Connell & Thompson, 1986) was employed. In such an experimental design, effects of treatment are assessed at regular intervals for each patient separately. In the present study, as treatment was extended to atypical or typical members of a superordinate category, generalization to the remaining examples was examined. The emergent naming patterns provided information regarding the re-organization and representation of semantic categories.

Prior to application of treatment, during the baseline phase, naming of all 48 examples of two categories (N = 24) was tested. Picture naming was then trained using selected examples of one superordinate category, with the order of categories and exemplar typicality counterbalanced across participants. During treatment, naming of all 24 examples in the category were assessed every second treatment session. These naming probes constituted the dependent variable in the study and naming accuracy over time was assessed. See table 1 for order of treatment for the four patients.

Criteria for acquisition of naming of trained items was 7/8 items named correctly on two consecutive naming probes. Generalization to naming of untrained examples was considered to have occurred when a 40% change over baseline levels was noted for untrained examples. If generalization to naming of untrained items was observed, treatment was shifted to the second category. If generalization to naming of untrained items was not observed, treatment was shifted to the next group (i.e., intermediate) within the same category.

Table 1: Order of treatment for the four participants

	P1	P2	P3	P4
Order of treatment				
1.	Birds 1.Typical 2. Inter 3.Atypical	Birds 1 Atypical 2. Inter 3.Typical	Vegetables 1.Typical 2. Inter 3.Atypical	Vegetables 1.Atypical 2. Inter 3.Typical
2.	Vegetables 1.Typical 2. Inter 3.Atypical	Vegetables 1.Atypical 2. Inter 3.Typical	Birds 1.Typical 2. Inter 3.Atypical	Birds 1.Atypical 2. Inter 3.Typical

Treatment

For each participant, one subset of items within a category (typical, intermediate or atypical) was trained at a time. In each treatment session, participants practiced the following steps for each of the eight examples of a subset: a) naming the picture, 2) sorting pictures of the target category (N=24) with three distracter categories (N=36), 3) identifying 6 semantic attributes applicable to the target example from a set of 35 features of the superordinate category, 4) answering 15 yes/no questions regarding the presence or absence of a set of semantic features about the target example. Distracters on this task included semantic features from the target category not applicable to the target, and features from unrelated superordinate categories.

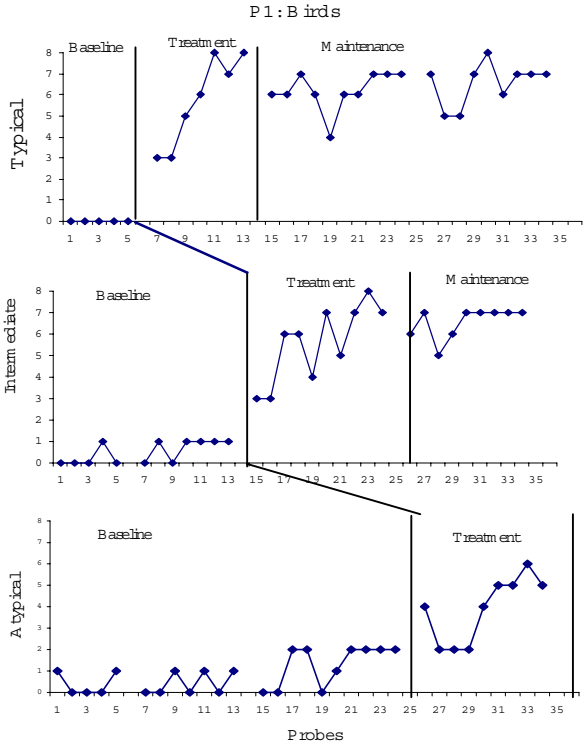


Figure 1. Naming accuracy on typical, intermediate and atypical items for the category birds for Participant 1

Results

Participant 1 Following five baseline sessions, treatment was initiated on typical items on the category birds. While naming of typical items improved to criterion (7/8 for two consecutive sessions), generalization to naming of intermediate or atypical examples was not observed. Treatment then was shifted to intermediate examples, following which improvement was observed on those items with no changes noted for atypical examples. Once criterion was achieved for intermediate examples, treatment was finally shifted to atypical examples and improvement was noted for the trained atypical items (see Figure 1). Administration of probes at phases denoting change of treatment set revealed no changes in items of vegetables.

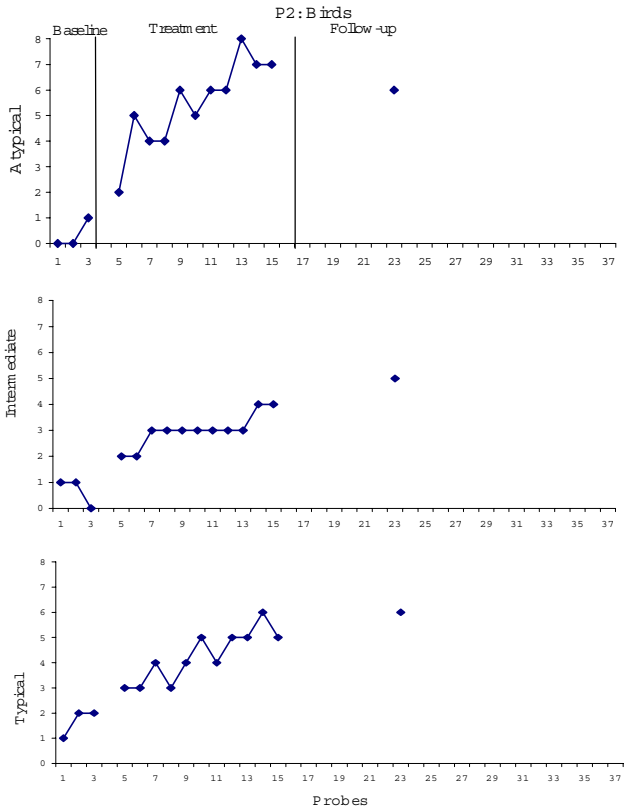


Figure 2. Naming accuracy on typical, intermediate and atypical items for the category birds for Participant 2

Participant 2 Following three baseline sessions, treatment was initiated for atypical examples of birds. Performance on naming of atypical examples improved to criterion (7/8 for two consecutive sessions), while generalization to naming of intermediate and atypical examples was noted (see Figure 2). Treatment then was shifted to vegetables. Following two baseline sessions, treatment was initiated on atypical examples of vegetables. Acquisition of atypical items for vegetables

was observed, and once again, generalization was noted for intermediate and typical examples, denoting replication within the participant across categories. Follow up probes administered within six weeks of completion of treatment indicated maintenance levels comparable to treatment levels.

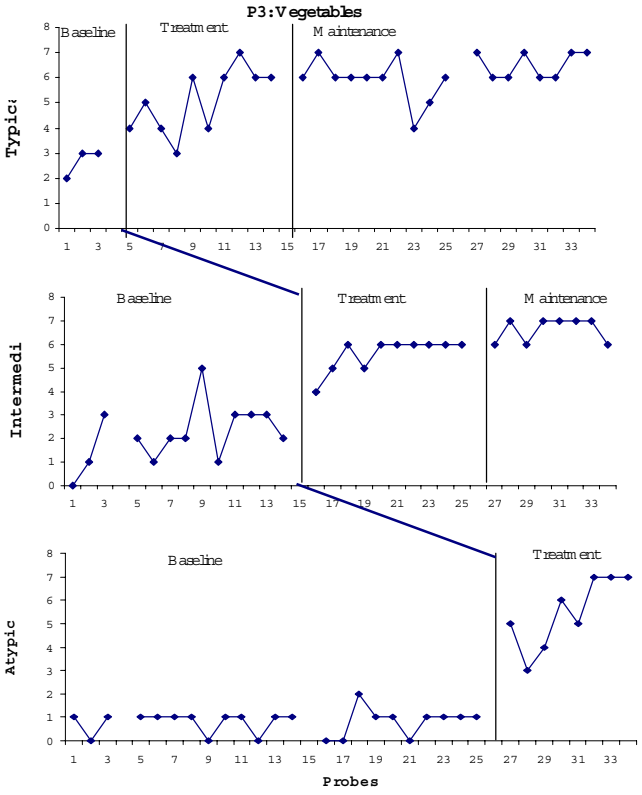


Figure 3. Naming accuracy on typical, intermediate and atypical items for the category vegetables for Participant 3

Participant 3 Following three baseline sessions, treatment was initiated on typical examples of vegetables. While an acquisition curve for typical items was discernible, criterion of 7/8 accuracy for typical examples was not achieved after 20 treatment sessions. Treatment was then shifted to intermediate examples, once again acquisition of trained items was noted but criterion was not achieved. Finally, treatment was shifted to atypical examples. Performance on those items reached criterion, while performance on typical and intermediate items was maintained (see Figure 3). Administration of probes at phase change revealed no changes in items of birds. For both participant 1 and participant 3, due to the extended duration, treatment was discontinued after completion of the first category.

Participant 4 Following five baseline sessions, treatment was initiated for atypical examples for vegetables. Performance on naming of atypical examples improved to criterion, with generalization noted on intermediate and atypical examples (see Figure 4) Treatment then was shifted to birds. Following two baseline sessions, treatment was initiated on atypical examples of birds. Acquisition of atypical items for birds was observed, while once again, generalization was noted for intermediate and typical example, once again providing a replication within participant across categories. Follow up probes administered within six weeks of completion of treatment indicated maintenance levels comparable to treatment levels.

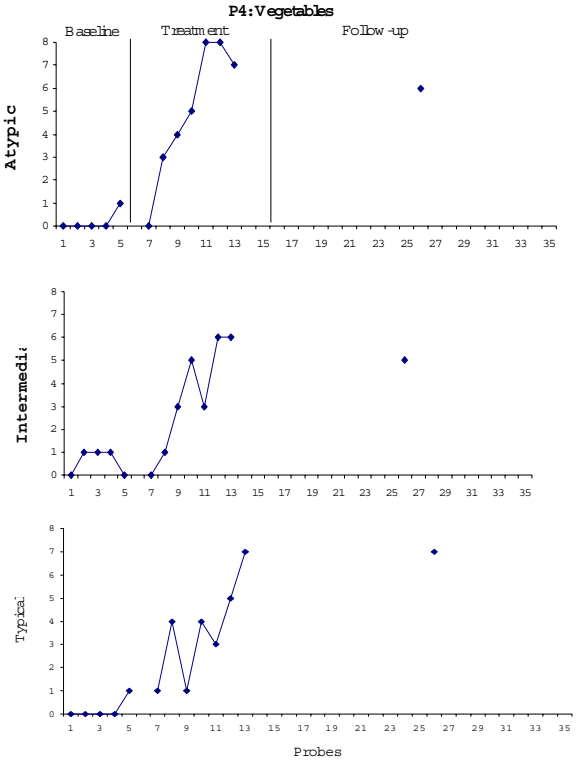


Figure 4. Naming accuracy on typical, intermediate and atypical items for the category vegetables for Participant 4

Discussion

The present experiment demonstrates that training atypical examples of a category and their semantic features results in generalization to naming of intermediate and typical examples of the category. This finding was observed in Participants 2 and 4 across two categories even when the order of categories was counterbalanced across the two participants. Training typical examples and their semantic features, however, did not result in generalization to the intermediate and atypical examples, as observed in Participant 1 and 3.

These findings suggest that because atypical examples are dissimilar to one another and to the category prototype, these examples collectively convey

more information in terms of semantic features about the variation that can occur within the category than do typical examples. Heightening access to featural information relevant to semantic categories, therefore facilitate access to more typical items within a category. While it has been demonstrated that greater coverage of a category's features can lead to stronger inductive generalizations (Sloman, 1993), current models of typicality do not explain the treatment effects observed in the present experiment. For instance, findings of the present experiment cannot be explained by the two-stage feature comparison model (Smith, Shoben, & Rips, 1974), since this model only explains category membership and not exemplar access.

Similarly, prototype/family resemblance models (Hampton, 1979; Rosch & Mervis, 1975) suggest that categories are represented by a set of weighted semantic features as a function of typicality, but do not explicitly state the relation between these summary feature representations and the phonological representations of specific examples. Moreover, prototype models do not specify how the various examples in a category are connected to each other, an element crucial to the explanation of the present experiment.

Exemplar models (e.g., Heit & Barsalou, 1996) come closest to explaining the results of the present experiment in that typical and atypical examples are represented as specific instances in the category that have been previously encountered. Therefore, it can be assumed that these specific representations are associated with their phonological representations. However, if examples of a category are represented as abstractions of specific instances, the exemplar models do not explain why training semantic features of atypical examples would result in improvements in naming of typical examples.

In summary, although all models of typicality explain possible differences in the representation of typical and atypical examples, they do not predict why training semantic features of atypical examples would improve phonological access of not just atypical examples but of intermediate and typical examples as well. More importantly, these models do not predict why training semantic features of typical examples would result in no improvement in the phonological access of intermediate and atypical examples. Even Plaut's connectionist model (1996), which motivates the present experiment, does not explain the mechanism involved in accessing improved phonological forms. This model describes a reading via meaning task with four layers, orthographic input layer, intermediate layer, a semantic layer and clean up layer. To generate semantic features, the prototype represents a set of semantic features (or binary values) with a high probability of becoming active. Typical examples share most of their features with the prototype, while atypical examples share few features with the prototype. Therefore, while this model provides an explicit account on the extent of difference between typical, atypical examples and the prototype, the nature of these features (whether defining or characteristic), and the nature of the examples (whether summary representations or specific instances) are also unclear.

Any explanation for the present experiment should therefore account for the following: (a) effects of treatment on improvements in semantic feature representation, (b) influence of strengthened semantic representations on access to phonological forms, (c) selective strengthening of connections between atypical and typical phonological representations (and not the other way around). A combination of interactive activation models (Dell, 1986; Stemberger, 1985) and prototype models of typicality provide such an explanation. Two levels of representation are hypothesized, semantic and phonological, and the connections between the semantic and phonological levels are bi-directional and excitatory while connections within each level are inhibitory. Within the semantic level, each example of a category (e.g., bird) is a summary representation of weighted semantic features, which interfaces with the lexical representation of the example. Items that are typical exert greater lateral inhibition on other examples within the category, due to their similarity with the category prototype. Less typical items exert less lateral inhibition on corresponding examples. This is because less typical items are dissimilar from the category prototype and illustrate the variation of semantic features that can exist (e.g., cannot fly, lives near water). Training semantic features of atypical examples strengthens their corresponding lexical representation and by the nature of the weak lateral inhibition, strengthens the representations of intermediate and typical examples as well. These strengthened semantic representations exert an excitatory influence on their corresponding phonological representations, which are raised above a resting threshold level. It is hypothesized that items directly trained receive a greater unit of activation to cross the resting threshold than untrained items.

Training typical examples on the other hand, only strengthens the semantic representations of the typical examples, and since these features convey no information about the variation of semantic features that can occur in the category, they have no influence on the semantic representations of intermediate and atypical examples. Therefore, the lateral inhibition exerted by the semantic representations of typical examples on intermediate and atypical examples does not reduce following treatment. Consequently, only the strengthened typical representations can successfully raise their corresponding phonological representations

above the resting threshold. The unchanged semantic representations for intermediate and atypical examples, can exert no excitatory influence on their corresponding phonological representations, and therefore have to be trained directly in treatment to be named successfully. These hypotheses are currently being tested using a connectionist network simulation.

Finally, results of the present experiment have significant implications for rehabilitation in aphasia. These results, although counter-intuitive to traditional treatment approaches, suggest that training naming of atypical examples is a more efficient method of improving naming items within a category than training typical items. Interestingly, training more complex items which encompass variables relevant to simpler items have been demonstrated in other language domains. Training complex syntactic structures results in generalization to simpler ones in agrammatic aphasic patients (Thompson, Ballard & Shapiro, 1998; Thompson et al., 1997) and training complex phonological forms results in improvements to simpler forms in children with phonological deficits (Gierut et al., 1996, 1999). These results also provide important insights into the mechanisms of relearning in patients with brain damage. In these individuals, it is assumed that language organization is fractionated following brain damage. The goal of language treatment is then to compensate and maximize the use of spared functions. The results of the present experiment suggest that relearning of category structure and corresponding phonological representations can be re-established in a more efficient way than previously thought.

References

Connell, P. K., & Thompson, C. K. (1986). Flexibility of single subject design. Journal of Speech and Hearing Disorders, 51, 214-225

Dell, G. S. (1986). A spreading activation theory of retrieval in sentence production. Psychological Review, 92, 283-321.

Gierut, J. A. (1999). Syllable onsets: clusters and adjuncts in acquisition. Journal of Speech Language and Hearing Research, 42 (3), 708-726.

Gierut, J. A., Morrisette, M. L., Hughes, M. T., Rowland, S. (1996). Phonological treatment efficacy and developmental norms. Language, Speech and Hearing Services in Schools, 27 (3), 215-230.

Goodglass, H., & Baker. E. (1976). Semantic field, naming and auditory comprehension in aphasia. Brain and Language, 10, 318-330.

Grober, E., Perecman, E., Kellar, L., & Brown, J. (1980). Lexical knowledge in anterior and posterior aphasics. Brain and Language, 10, 318-330.

Grossman, M. (1981). A bird is a bird: Making references within and without superordinate categories. Brain and Language, 12, 313-331.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. Journal of Verbal Learning and Verbal Behavior, 18, 441-461.

Heit, E., & Barsalou, L. (1996). The instantiation principle in natural language categories. Memory, 4, 413-451.

Larochelle, S., & Pineu, H. (1994). Determinants of response times in the semantic verification task. Journal of Memory and Language, 33, 796-823.

McNeil, M. R., Doyle, P. J., Spencer, K., Goda, A. J., Flores, D., & Small, S. L. (1998). Effects of training multiple form classes on acquisition, generalization and maintenance of word retrieval in a single subject. Aphasiology, 12 (7-8), 575-585.

Plaut, D. C. (1996). Relearning after damage in connectionists networks: Toward a theory of rehabilitation. Brain and Language, 52, 25-82.

Pring, T., Hamilton, A., Harwood, A., & McBride, L. (1993). Generalization of naming after picture/word matching tasks: only items appearing in therapy benefit. Aphasiology, 7 (4), 383-394.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic distance. Journal of Verbal Learning and Verbal Behavior, 12, 1-20.

Rosch, E. (1975). Cognitive representation of semantic categories. Journal of Experimental Psychology: General, 104, 192-233.

Rosch, E., & Mervis, C. (1975). Family resemblances: Studies in internal structure of categories. Cognitive Psychology, 7, 573-604.

Sloman, Steven A. (1993) Feature-based induction. Cognitive Psychology, 25 (2), 231-280

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model of semantic association. Psychological Review, 81, 214-241.

Stemberger, J. P. (1985). An interactive model of language production. In A. W. Ellis (Ed.) Progress in the psychology of language Hillsdale, NJ: Lawrence Erlbaum.

Thompson, C. K, Shapiro, L., Ballard, K. J., Jacobs, J. J. Schneider, S. L., & Tait, M. E. (1997). Training and generalized production of wh- and NP movement structures in agrammatic speakers. Journal of Speech, Language and Hearing Research, 40, 228-244.

Thompson, C. K., Ballard, K., & Shapiro, L. (1998). Role of syntactic complexity in training wh-movements structures in agrammatic aphasia: Order for promoting generalization. Journal of International Neuropsychological Society, 4, 661-674.

Visual Statistical Learning in Infants

Natasha Zoe Kirkham (nzk2@cornell.edu)
Department of Psychology, Cornell University
Ithaca, NY 14853 USA

Jonathan Andrew Slemmer (jas234@cornell.edu)
Department of Psychology, Cornell University
Ithaca, NY 14853 USA

Scott P. Johnson (sj75@cornell.edu)
Department of Psychology, Cornell University
Ithaca, NY 14853 USA

Abstract

Statistical probability theory posits that we learn about regularly-occurring events in the perceptual environment by determining the likelihood of each event's occurrence (Aslin, Saffran, & Newport, 1998). The current study investigates infants' ability to extract properties of repetitive visual events and represent predictable combinations of visual elements. Using a novelty-preference paradigm, 2-, 5-, and 8-month-old infants were habituated to a continuous stream of colored shapes that were presented in a statistically predictable pattern, and then tested alternatively on the same sequence and a randomly-ordered sequence. The randomly-ordered sequence differed from the originally presented sequence only in between-shape transitional probabilities. At each age, infants demonstrated a significant novelty preference for the random sequence. In conjunction with Marcus, Vijayan, Rao, and Vishton (1999) and Saffran, Aslin, and Newport's (1996) work looking at statistical learning in language with 7- and 8-month-olds, these results can be taken as preliminary evidence of a domain general learning mechanism.

Introduction

One of the fundamental questions asked by developmental psychologists concerns how infants learn so much in so little time, with apparently very little explicit instruction. Research suggests that, as adults, we are remarkably good at implicit learning (e.g., see Stadler & Frensch, 1998 for a review of the implicit learning literature). Implicit learning can be defined as non-conscious facilitation of task performance due to information acquired during previous exposure. Given the robust nature of implicit learning skills in adults, it is perhaps a reasonable assumption that these skills play some role in early learning. The research on implicit learning in children suggests that they too show implicit sequence learning to the same degree as adults (Meulemans, Van der Linden, & Perruchet, 1998; Thomas, 1998). In other words, children showed increased reaction time in a task that contained a predictable sequence, and, like adults, did not have explicit knowledge of this sequence. There were no reports in the literature, however, with participants younger than 3 years of age until recent studies of statistical learning, a form of implicit learning based on statistical regularities in the perceptual environment.

Saffran, Aslin, and Newport (1996) and Alsin, Saffran, and Newport (1998) presented evidence that 8-month-old infants determine the statistical probability of neighboring speech sounds based on a 2-minute exposure. Infants heard four three-syllable "words" composed of 12 unique syllables (e.g., *tupiro*, *golabu*, *dapiku*, and *tilado*), presented in a continuous stream in random order (e.g., *dapikutupirotiladogolabutupiro ...*). Between-word spaces were removed, as were all other cues to word boundaries (e.g., rhythm, intonation, and stress). Thus, the only cues to word boundaries were the transitional probabilities between syllable pairs. For example, the transitional probability of *tu-pi* in this corpus is 1.00, because *pi* always follows *tu* within the word *tupiro*, whereas the probability of *ro-go* is .33, because *golabu* is one of three words that can follow *tupiro*.

After exposure, Saffran et al. (1996) presented infants with both familiar words from the corpus and "nonwords." Nonwords were created by combining the last syllable of one word with the first two syllables of a second word (e.g., *rogola* and *butupi*). Infants showed greater interest in the nonwords than in the words. On the logic that infants often exhibit a post-familiarization novelty preference (Bornstein, 1985), these results suggest that they detected the difference between words and nonwords. This outcome is necessarily based on learning of the transitional probabilities defining the stimuli.

This finding is consistent with a powerful statistical learning mechanism that supports language acquisition, and gives rise to questions concerning the generality of such mechanisms. It is possible, for example, that other kinds of knowledge are gained during infancy by learning statistical regularities in the environment. Indeed, Saffran, Johnson, Aslin, and Newport (1999) presented evidence that both adults and 8-month-old infants can perform the same statistics when presented with "words" consisting of non-linguistic tone sequences. These data suggest that statistical learning is not just a linguistic mechanism, allowing us to parse words from noise. If this is the case, then perhaps statistical learning is a mechanism that bolsters all sorts of learning in many different domains. Perhaps statistical learning is a domain general mechanism.

Following from this hypothesis the next logical step is to look at statistical learning in a non-auditory domain. The visual domain provides a lot of opportunities for patterns to present themselves, and the current study tested infants' ability to pick out statistical regularities in a stream of visual events. The data present evidence that 2-, 5-, and 8-month-old infants are able to extract properties of repetitive visual events in a way that allows them to represent a predictable combination of visual elements. Results from this series of studies showed that at each age, infants demonstrated a significant novelty preference for a visual sequence that differed from the originally presented sequence only in between-stimulus transitional probabilities.

Method

Participants

A total of 48 infants participated in this study: 16 infants at 2 months of age ($\underline{M}$ = 1.95 months), 16 infants at 5 months of age ($\underline{M}$ = 5.10 months) and 16 infants at 8 months of age ($\underline{M}$ = 7.99 months). Infants were recruited through a database of infants in the Ithaca, NY area. Informed consent was obtained from all parents, and the infants received a small toy or T-shirt as thanks for participation. All infants were full term and healthy. In addition to the 48 infants included in data analyses an additional 16 infants were tested but were not useable. Eight of these infants were 2-month-olds who either fell asleep during testing or were so fussy that looking times could not be judged correctly. In the 5- and 8-month-old group, data from one infant was not included due to equipment failure; the remaining seven infants were not included because they were so fussy that testing had to be terminated prior to the presentation of test trials.

Stimuli and Apparatus

Stimuli consisted of six looming colored-shapes (pink diamond, red octagon, yellow circle, blue cross, turquoise square, and green triangle) presented on a 53 cm computer monitor. The six colored-shapes were vector shapes that loomed from 2.35˚ to 14.59˚ of visual angle. Each stimulus loomed from 4 cm (2.35˚ visual angle) to 24 cm (14.59˚) in 1000 ms. There was no pause between stimulus presentations. The stimuli appeared in a continuous stream of randomly-ordered pairs (e.g., Pair 1: turquoise square followed by blue cross; Pair 2: yellow circle followed by pink diamond, Pair 3: green triangle followed by red octagon), with only transitional probabilities defining between-stimulus boundaries (see Figure 1 for an example of one shape sequence). The transitional probability within pairs was 1.0 and between pairs was 0.33 (see Figure 2 for an example of the transitions). In other words, for an individual infant, the pairs were always the same, but the order of the pairs within the sequence was random (e.g., turquoise square, blue cross, yellow circle, pink diamond, green triangle, red octagon, yellow circle, pink diamond, green triangle, red octagon, turquoise square, blue cross, turquoise square, blue cross….)

Procedure

Infants were tested individually and sat on a parent's lap 93 cm from the television monitor. The parent was instructed not to pay attention to their baby or to watch the screen. Infants were first shown the original stimulus sequence until looking declined according to a preset habituation criterion, determined by a sliding window algorithm that calculated over the course of a block of four trials when the infant's looking time had decreased by 50% from baseline. After habituation, the infants viewed six test displays alternating between familiar sequences, composed of the same three colored shape pairs, and novel sequences, produced by random recombinations of the same colored-shapes. The only difference, therefore, between the familiar and the novel sequences was that in the familiar sequences the first member of a colored shape pair predicted the second member, whereas in the novel sequences the colored shapes had no predictive value. There were six test trials in total (three familiar, and three novel). Test trials were counterbalanced across infants so that half the infants saw a familiar trial first and half the infants saw a novel trial first. The actual structure of the pairs was randomized across infants, so that it was very unlikely that any two infants saw the exact same pair sequence.

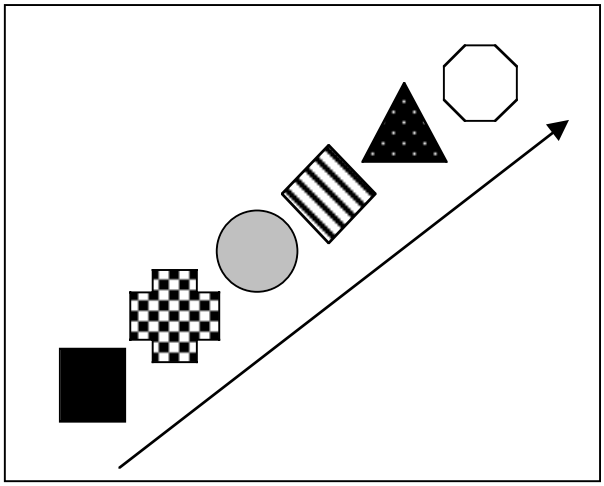


Figure 1: Example of a Shape Sequence (the actual shapes had unique colors, not black and white patterns).

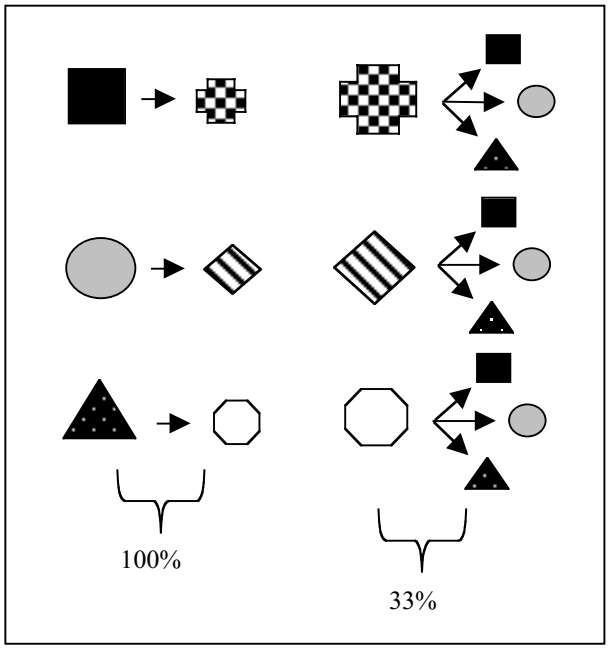


Figure 2: Example of the transitional probabilities between shapes.

Results

We hypothesized that infants would look longer at the novel sequences, if they were able to extract the visual statistical regularities available in the habituation stimuli.

The dependent measure was looking time at the familiar vs. the novel sequences. Looking time data in some cells were positively skewed (i.e., there were a few extremely long looking times); therefore, all data were log-transformed prior to analysis. A 3 (age group: 2-, 5-, or 8-month-olds) x 2 (test display: habituated vs. random sequence) ANOVA yielded a significant main effect of age, $F(2, 45) = 26.19$, $p < 001$, the result of longer looking overall by the youngest infants (see Table 1). As was predicted, there was a reliable main effect of test display, $F(1, 45) = 17.89$, $p < .001$, the result of longer looking at the random test display (see Table 1). Planned comparisons showed that at each age infants looked longer at the random test display than at the familiar test display (at 2 months of age, $F(1, 45) = 5.47$, $p = .024$; at 5 months of age, , $F(1, 45) = 6.68$, $p = .013$; at 8 months of age, $F(1, 45) = 5.77$, $p= .02$; see Figure 3).

Table 1: Mean looking time in sec per test display according to age group.

Age Group	Familiar Sequence	Novel Sequence
2 months	23.67	32.95
5 months	8.10	11.39
8 months	6.07	9.28

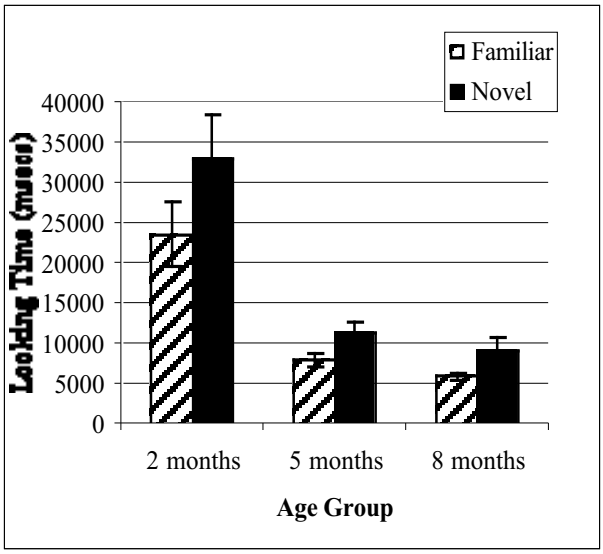


Figure 3: Infants' Looking Time to Familiar and Novel Patterns

Discussion

These results reveal that even very young infants are capable of picking up on statistical regularities in the visual environment. Moreover, they are discriminating between two visual sequences that differ *only* in these statistical probabilities. The sequences offered no cues as to the pairings: there were no pauses between stimulus presentations, the presentations were the same length, and the colored-shapes were not salient objects. Yet, each age group showed significantly longer

looking time to the novel sequences. This evidence that statistical learning is evident in young infants, in the visual domain, lends support to the hypothesis that this could be an early developing mechanism that supports learning across domains.

It is, of course, worth noting that this is only the first step in addressing the ways in which learning develops. We have yet to determine exactly how infants are encoding the information. Is it pure statistics? Are babies taking note of co-occurrences of events and stimuli, and judging the probabilities of those co-occurrences or are they abstracting these statistics into higher-order rules? Research on language learning has suggested that developmental constraints on learning actually provide the necessary prerequisites for later complex learning (Elman, 1993; Newport, 1988, 1990). Our research offers the same suggestion; early learning in the visual domain starts with the acquisition of the simplest patterns, and over development more complex patterns are acquired. This domain general learning could provide the beginnings of domain specific learning: as learning becomes more complex, the types of learning differentiate according to domain

As adults, we are brilliant implicit learners, and this learning has very little decay and requires very little effort. As infants, all we do is look, listen and touch; we absorb information from the environment and quickly pick up on the causal relationships we experience around us. For example, it does not take an infant very long to figure out that crying creates a desired effect, that of a parent's presence, and that the presence of a parent tends to predict food or comfort. This type of associative learning seems natural and adaptive. But, what if all initial learning is associative, and not dependent on a conditioned reward/punishment outcome. Note that the point here is to focus on initial learning, not to suggest that all learning develops in the same way. Perhaps the associations are all that is needed to elicit a type of implicit learning. Admittedly, the associations that are relevant and salient and do produce pleasant or unpleasant end results might be learned faster and remembered longer. We do not think that the infants we tested are going to predict the arrival of a turquoise square every time they see a yellow circle, for example, but it is interesting to observe that the associations were there at least long enough for the babies to notice a difference when the turquoise square did not predict a yellow circle. What is most interesting about the statistical learning process is not that it may be domain general but that it seems to work in situations that do not have reward/punishment end results, and therefore, seems capable of supporting a great deal of initial knowledge acquisition.

Acknowledgments

The authors would like to thank the many parents and infants for their participation, and all the members of the Cornell Baby Lab for help with infant recruitment. As well, we would like to thank Richard Aslin for many helpful comments. The research was supported by NSF grant BCS-9910779.

References

Aslin, R. N.; Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological-Science, 9,* 321-324.

Bornstein, M. H. (1985). Habituation of attention as a measure of visual information processing in human infants: Summary, systematization, and synthesis. In G. Gottlieb & N. A. Krasnegor (Eds.), *Measurement of audition and vision in the first year of postnatal life: A methodological overview*. Norwood, NJ: Ablex.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48,* 71-99.

Marcus, G.F., Vijayan, S, Rao, S.B., & Vishton, P.M. (1999). Rule learning by seven-month-old infants. *Science*, *283(*5398), 77-80.

Meulemans, T, Van der Linden, M., & Perrouchet, P. (1998). Implicit learning in children. *Journal of Experimental Child Psychology, 69(*3), 199-221.

Newport, E. L. (1988). Constraints on learning and their role in language acquisition: Studies of the acquisition of American Sign Language. *Language Sciences, 10,* 147-172.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science, 14,* 11-28.

Saffran. J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274,* 1926-1928.

Saffran, J.R, Johnson, E. K., Aslin, R.N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70(*1), 27-52.

Spelke, E. S., & Van de Walle, G. (1993). Perceiving and reasoning about objects: Insights from infants. In N. Eilan, R. A. McCarthy, & B. Brewer (Eds.), *Spatial representation: Problems in philosophy and psychology*. Oxford: Blackwell.

Stadler, M.A., & Frensch, P.A. (Eds.) (1998). *Handbook of implicit learning.* Thousand Oaks: Sage Publications

Thomas, K.M. (1998). Behavioral and electrophysiologic measures of implicit learning in children. *Unpublished doctoral dissertation.*

Episode Blending as Result of Analogical Problem Solving

Boicho Kokinov (bkokinov@nbu.bg)[12]
Neda Zareva-Toncheva (nzareva@cogs.nbu.bg)[2]
[1]Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. G. Bonchev Street, bl.8
Sofia 1113, Bulgaria

[2]Central and East European Center for Cognitive Science, Department of Cognitive Science and Psychology,
New Bulgarian University, 21 Montevideo Street
Sofia 1635, Bulgaria

Abstract

We know that misinformation presented in interrogating questions or in advertising produces blendings, that even imagining a possible episode might produce blending as well, however, we do not know whether reasoning and problem solving can produce the same effect. On the other hand, models of analogy-making assume "perfect memory" for old episodes. The AMBR model of analogical problem solving has mechanisms for interaction between memory and reasoning which explain partial memory and memory distortions and has predicted blending effects which are due to the reasoning process. Such predictions have no parallel in any other model we know of. There has been no experimental support for these predictions so far. The current paper describes an experiment explicitly designed to test these predictions. It consists of three sessions: 1) solving three problems, 2) solving two more target problems by analogy with some of the problems in the first session, and 3) reproduction of the three problems in the first session. The results demonstrate that the degree of blending in the recalled stories depends on the target problem solved in the second session.

Motivation

There is a considerable amount of research demonstrating various types of memory distortions, such as schematisation, blending, and false memory illusions (Bartlett, 1932; Loftus, 1977; Loftus, 1979; Loftus, Feldman, & Dashiell, 1995; Loftus, Miller, & Burns, 1978; Loftus & Palmer, 1974; Moscovitch, 1995; Neisser, 1998; Nystrom & McClelland, 1992; Reinitz, Lammers, & Cochran, 1992; Schacter, 1995). These findings, however, have been established in "pure memory tasks" conditions. Very rarely researchers have tried to integrate memory tasks with other kinds of cognitive tasks in order to explore whether there will be interactions between them. Thus for example, we know that misinformation presented in interrogating questions or in advertising produces blendings, that even imagining a possible episode might produce blending as well, however, we do not know whether reasoning and problem solving can produce the same effect.

At the theoretical end these findings are typically explained by the constructive nature of human memory, however, very few models exist that lay out the specific memory mechanisms that could explain the memory distortion effects. The key concept has always been the postulation of distributed representations of some type in human memory. These include Hintzman's (1988) multi-trace model, Metcalfe's (1990) CHARM model, and McClelland's (1995) PDP-type of model. The last two models explicitly deal with memory blending effects.

On the other hand, models of human analogical reasoning which necessarily include a memory component (since they have to explain how analogous episodes are retrieved) simply ignore both the experimental findings about memory distortions and the theoretical ideas about the constructive nature of human memory. They typically assume "perfect" memory for past episodes and even some kind of nice organization of memory that would allow the retrieval of the relevant episode (Forbus, Gentner, & Law, 1995; Hummel & Holyoak, 1997; Kolodner, 1984; Thagard, Holyoak, Nelson, & Gochfeld, 1990; Warton, Holyoak, & Lange, 1996). All these models assume centralized representation of episodes which means that episodes are either retrieved as a whole (and than mapped onto the target problem description) or they fail to be retrieved. No blending of episodes may occur, no false memories can arise, no partial retrieval can happen. Surprisingly, this is true even for the LISA model (Hummel & Holyoak, 1997) which is based on distributed representations, since the representations are truly distributed only in working memory, but the episode representation in LTM is highly centralized, although distributed in some sense (in the same sense in which the representation is distributed in the model described in the current paper). A list of all units representing a single episode is assumed and if the episode wins the competition between episodes then all it corresponding units are switched from dormant to active state.

The main motivation for recent developments in the AMBR model of analogical problem solving (Kokinov, 1998; Kokinov & Petrov, 2000, 2001; Petrov & Kokinov,

1998, 1999) was to propose mechanisms for interaction between memory and reasoning which will allow for explaining memory distortions and blendings among other things. AMBR has predicted blending effects which are due to the reasoning process. Such predictions have no parallel in any other model we know of. However, we do not know of any experimental support for them either. That is why an experiment was designed to test these predictions. The current paper describes this experiment and its outcomes.

AMBR Model and its Predictions

The AMBR model of analogical problem solving was introduced in (Kokinov, 1988) and then further developed over the years (Kokinov, 1994a, Kokinov, 1998, Kokinov & Petrov, 2001). It is based on a general cognitive architecture called DUAL (Kokinov, 1994b,c) which relies on emergent computations produced by a society of micro-agents (Minsky, 1986; Hofstadter, 1995). The latest developments (Kokinov, 1998, Kokinov & Petrov, 2000, 2001; Petrov & Kokinov, 1998, 1999) are directed towards building a more psychologically plausible memory for episodes and thus allowing for memory distortions to take place. Moreover, a prediction is made that the reasoning process during the analogical problem solving may itself produce memory distortions.

Episode Representation

Episode representation in AMBR is highly decentralized, which means that each episode is represented by a large coalition of micro-agents each of them representing some aspect of the situation. We may call this representation "distributed at a higher level" by analogy to the connectionist representations which are distributed over a set of simple features. In reasoning, and in analogy-making in particular, the relations are even more important than single features and therefore they have to be explicitly represented. The agents here are more complex than the connectionist units (Kokinov, 1994a, b, c) and roughly speaking they represent a whole proposition (such as "there is a coffee pot", "the coffee-pot is made of metal", "the coffee-pot is on the plate", "the plate is hot", "the water is in the coffee-pot", "proposition 1 causes proposition 2", etc.). Thus the episode is represented by a set of such propositions (by a coalition of the corresponding micro-agents), but there is nowhere in memory a list of all propositions (all micro-agents) involved in a given episode. That is why we call this representation decentralized. There is one agent which refers to the unique point in time and space – when and where the event has happened, and all agents from a given coalition have links directed to it. In this way the system can differentiate among propositions belonging to different episodes, however, there are no links from this agent to the agents in the coalition, i.e. it does not know any of the members of the coalition. Therefore this agent cannot contribute to the retrieval or construction process.

Each agent has a level of activation that changes dynamically. The activation level determines the degree to which the information contained in that agent is available and the degree to which this agent participates in the computational process (Kokinov, 1994b,c, Petrov & Kokinov, 1999). Working memory is considered to be the active part of long term memory, i.e. the set of all active agents at a given moment. There is a process of spreading activation as well as a process of decay which guarantees that an agent that do not receive activation will soon quit WM.

Episode "Retrieval" or "Construction"

Episode "retrieval" corresponds to the process of activation of the agents representing various aspects of the event and bringing them into WM. This means that typically the recall is only partial since there is no way to guarantee the activation of all members of a coalition. Some coalitions will be stronger (having stronger links between the agents) and therefore the corresponding episode will tend to be reproduced more fully, other coalitions are weaker and only few aspects of the episode are reproduced.

However, many agents belonging to other coalitions will also turn out to be activated – agents representing some general knowledge (concepts, facts, rules, etc.), agents representing aspects of other episodes. Thus the set of agents happened to be in WM will produce a description of the episode which is partial, but also containing intrusions from general knowledge and other episodes. Intrusions from other episodes are in fact blendings between two or more episodes. In fact, the representation of an old episode is not just retrieved from LTM, but is actively constructed in WM. The process of spreading activation is an automatic one but it depends on the current state of WM, the goals of the system and its input from perception. Thus the reasoning process, which runs in parallel to the memory process, interferes with this process of episode construction and in fact even guides it to a certain extend.

Mapping Guidance in Episode Construction

In the context of analogical problem solving the construction of the old "retrieved" episode is guided by the mapping process between this episode and the current target problem. The mapping process in AMBR does not start after the old episode is retrieved as in all other models of analogy-making, but runs in parallel to it. This makes it possible the already established partial mapping to guide the episode construction in such a way that the old episode is reconstructed in directions which allow better alignment between base and target. For example, the intrusions from general knowledge and from other episodes will be not arbitrary, but will correspond to elements of the target description which do not have corresponding elements in the base description (they are either missing in the encoding of the episode or are simply not activated at the moment). The precise mechanisms for this episode extension are described elsewhere (Kokinov & Petrov, 2000). A simulation experiment with AMBR has demonstrated that

the parallel run of reasoning and "retrieval" in this model yields the retrieval of structurally similar episodes that would otherwise be not retrieved (Petrov & Kokinov, 1998).

AMBR's Prediction

AMBR's prediction relevant to the current paper is that intrusions from other episodes will happen more often if the currently constructed representation of the most active episode is missing elements which are important for the mapping process with the target problem. This emphasizes not any missing element, but elements which are crucial for the mapping. This is in contrast to a model that is very similar in flavor (Nystrom & McClelland, 1992; McClelland, 1995) which, however, is not sensitive to the structure but fills in any missing information. The latter model explains data from a memory experiment on sentence recall and is not intended to explore the relation between memory and reasoning. Structure might be unimportant in this case. In contrast, AMBR makes stronger predictions about the relevance and structural consistency of the intruded elements.

In this case the analogy-making process itself produces blending under certain conditions. Therefore the more partial the mapping between the target and the base problem is the more intrusions will occur and thus higher degree of blending between episodes will be observed. This is the prediction tested in the following experiment.

Experiment

The main idea of the experiment is the following one. Ask the participants to solve several base problems and as a side effect to hold them in their long-term memory. Half of the participants will then solve one target problem and the other half another target problem. After that we will ask the participants to retell us the base problems as accurate and complete as possible. We will measure the degree of blending between the problems and expect that it will depend on which target problem has been solved.

Method

Design

The experiment consists of three sessions:

- Session 1: solving three base problems (A, B, C);
- Session 2: solving one of two target problems (T1 – partially analogous to A and partially analogous to B, or T2 – partially analogous to B and partially analogous to C);
- Session 3: recalling the problems from session 1.

The whole trick is that target problem T1 partially maps to both A and B, while target problem T2 partially maps to both B and C. That is, in order to solve problem T1 one needs to make a double analogy (with A and B) and use two different principles which are provided in A and B respectively. This requires that A and B are both partially "retrieved" in WM and partially mapped to T1. We may describe this situation as blending the two episodes A and B and constructing a new "old" episode AB that is then remembered. Thus later on in session 3 we would expect that participants who solved problem T1 will tend to blend episodes A and B. Reversibly, participants who solved target problem T2 in the second session will tend to produce more blends of B and C in session 3.

Thus we use a between group design. The independent variable is the type of target problem solved in session 2 and its relation to the base problems. The dependent variables have to measure the blending occurring in the retold stories in session 3. We use two types of measures:

- a binary variable – "yes/no" expert judgments of blended memories;
- degree of blending (a value between 0 and 1) – measured as the degree of mixture between statements related to each of the base problems A, B, and C.

Material

The problems used in the experiment both in session 1 and in session 2 have been designed to fulfill several criteria:

- they should be solved by some general principles that can then be applied to another problem;
- if the principles used for solving the base problems A, B, and C in session 1 are called PA, PB, and PC respectively, then the target problems in session 2 should be solved by a combination of two principles (T1 by PA and PB, and T2 by PB and PC).

In *problem A* we used a criminal story about an attempt to come in for money of a wrong person who actually killed the legatee and dressed and acted like her. She imitated successfully even the gestures of the dead relative, including her habit to arch her right eyebrow when asking questions. She practiced these gestures for a long period of time in front of a mirror. Finally, however, she was recognized as a fake legatee. The question is how she was recognized. And the correct solution relies on principle PA: "Left and right are reversed in an mirror image". Thus, the lady arched her wrong eyebrow.

Problem B was an expanded version of the radiation problem. The principle underlying the correct solution was PB: the convergence of several weaker X-rays in one point form a stronger X-ray.

Problem C involved baking 3 flat loafs in a small baking tin which can hold only 2 loafs. The question was what is the minimal time period required to bake the 3 loafs turning each of them on both sides. And the principle underlying the correct solution was PC which outlines a turning schema: you first bake one side of two loafs, then you turn one of them and replace the second loaf with the third one, and finally you bake the remaining sides of the second and the third loafs.

All three problems were told as folk tales with some superficial similarities in the plot (having kings, princesses, wise man, etc.), still they were quite different.

Target problem T1 was about a five-headed dragon that has to be killed, but he can be killed only if at the very same moment his 2nd head from the left to right is cut off and his heart is destroyed by a strong laser beam. However, there were several obstacles: you cannot look at the eyes of the

dragon directly because you will become blind, and also there were only 3 weak laser beams available. Thus the participants had to apply principle PA and to cut off the 2nd right head (instead of the 2nd left one) staying backwards and looking at the dragon into a mirror and principle PB to use a converging configuration of three weak laser beams.

Target problem T2 involved killing another dragon where again his heart should be destroyed by a strong laser beam and you have only three small ones, however, here the dragon was behind a moat full of lava which could be passed only by using three magic stones. The obstacle is that you can use each stone only once to step on each of its two sides. The solution involved principle PB from above, and principle PC used in problem C – the particular scheme of turning the stones and loafs respectively.

Procedure

During the first session the participants were told that they will undergo a series of experiments on human thinking and they will have to solve various problems. The problems were given one by one without an explicit time restriction. After the participants produced a written solution the experimenter read it aloud and if this was not the targeted solution she encouraged them to find an alternative solution, if this did not help a hint was given, and finally if the target solution was not found it was provided by the experimenter. The aim was all participants in session 1 to solve the base problems correctly and to acquire the basic principles PA, PB, and PC.

The second session followed after a period of 3 to 7 days. During this second session the group was split and half of the subjects solved problem T1, and the other half – T2. The second session was run again individually and the thinking aloud method was used, the speech of the participants was recorded. No hints were provided here.

The third session followed immediately after the second session. During this session the participants were asked to retell as accurate and complete as possible the problems from the first session. The stories were reproduced orally and tape-recorded.

Participants

48 undergraduate students participated in the experiment, but only 33 went through all sessions. 16 were female and 17 – male.

Results and Discussion

The records were transcribed and the protocols of the third session were used as the main data set. Each story was segmented into short phrases which express independent and understandable statements. The texts of the problems A, B, and C were also segmented into separate statements and their appearance in the body of the protocol was encoded. For example, we separated the text of problem C into 22 statements – C1-C22 and whenever a phrase (or its semantic equivalent) occurred in the narration of the subject the corresponding Ck was inserted in the protocol encoding.

Quite often when reproducing one of the problems participants inserted statements from one of the other two problems. This was exactly what we were counting. Thus for measuring the AB blending (blending between problem A and B) we counted how many As and how many Bs we have in the reproduction. The degree of blending was calculated as the ratio: number of As over number of Bs (when the As are less than the Bs), or reverse – number of Bs over number of As (when the As are more than the Bs). To put it differently we measured the percentage of intrusions in the text arising from another problem. Thus if the number of As or Bs are zero than no blending has occurred (degree of blending is 0), and when the number of As and Bs are equal then an absolute blend has been produced (with degree of blending equal to 1). The results are shown in Table 1 and Figure 1. As we can see the results are coherent with our hypothesis: we have higher degree of blending of type AB in group 1 (solving T1 in the second session) and higher degree of blending of type BC in group 2 (solving T2 in the second session). Just to remind that T1 required double analogy with A and B, and T2 required a double analogy with B and C. At the same time there is no difference whatsoever in the degree of blending of type AC which should be expected since none of the target problems required combining base A and base C. The performed analysis of variance showed a significant 2-way interaction between the groups (target problems solved) and the type of blending occurred ($F(2,62)=4.41$, $p<0.016$). The difference in the degree of AB blending is significant, but the difference in the degree of BC blending is not significant. Analyzing our data we found out that very few of the participants in group 2 were able to solve target problem T2 (in fact only 4 out of 16) and therefore they have not done the double analogical mapping with B and C. Thus we cannot expect the blending effect in this case. In contrast, problem T1 was solved by 11 out of the 17 participants. Evidently the second target problem was too difficult. We looked into the solved/unsolved difference and found support for this interpretation. The mean degree of BC blending in group 2 is: 0.148 for the subjects who solved T2 and 0.025 for the subjects who did not solved it, i.e. it is 6 times higher for the subjects who solved the target problem T2. The difference is significant at p-level 0.028 (measured by one-tailed t test, $t=2.45$)

Table 1: Mean degrees of blending in each group.

	# of subjects	AB blending	BC blending	AC blending
group 1 (solving T1~AB)	17	0.102	0.043	0.036
group 2 (solving T2~BC)	16	0.022	0.056	0.035

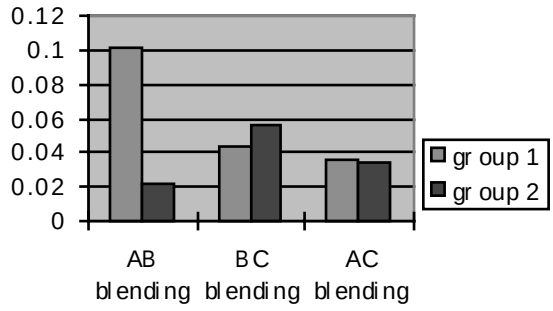


Figure 1: Degree of blending between the base problems reproduction depending on experimental group. There is a significant 2-way interaction between groups and type of blending.

Since we used a very formal method of measuring blending – the number of intrusions as registered in the protocols – we were curious to compare this to a more qualitative judgment done by human experts who may recognize whether there is a real blending or just some general knowledge intrusions or superficial mixture. Two independent judges had to read each protocol (without knowing neither about our hypothesis nor which group this protocol comes from). The experts had to judge whether there was a blending between some old problems. There was a high degree of agreement between the experts (about 10% disagreement where a third expert was called for judgment). The frequencies of blending of type AB, BC, and AC are presented in Table 2 and their percentage in Figure 2.

Table 2: Number of blendings as judged by experts.

	# of subjects	AB blending	BC blending	AC blending
group 1 (solving T1~AB)	17	11	6	5
group 2 (solving T2~BC)	16	5	8	5

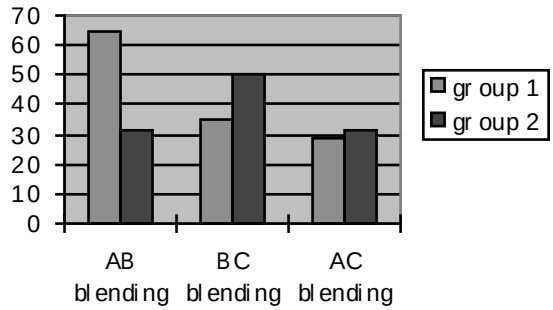


Figure 2: Percentage of blendings as judged by experts.

The results from expert judgments are generally coherent with the measurements of the degree of blending and again we have more AB blends in group 1, and more BC blends in group 2. The number of AC blends is almost equal in both groups. The log-linear analysis declared all the differences insignificant, however.

Conclusions

AMBR mechanisms of memory access and mapping work in parallel and interact with each other and thus predict mutual influence between these processes. One specific prediction is that when the target problem maps only partially with two different bases and a double analogy is required to solve it, then both bases are partially activated and elements from both episodes are brought into WM. In that way a blend between the two old episodes is produced and remembered. That is why a higher degree of blending is expected after subjects having solved such targets requiring double analogies.

This prediction has been experimentally tested in a series of three sessions in which the participants had first to learn the bases (by solving the problems in session 1), then to solve one additional target problem (session 2), and finally to retell the base stories (session 3). It turns out that higher degree of blending between two episodes is observed in the group where the target problem in session 2 required usage of a double analogy with these two episodes.

We consider these results as a first step in a series of experiments which have to test this hypothesis. We are currently running several more experiments varying the material, the design of the experiment and the timing of the sessions.

Acknowledgments

We would like to thank the AMBR research team for their continuous support and stimulating environment.

References

Bartlett, F. (1932). *Remembering*. Cambridge: Cambridge University Press.

Forbus K., Gentner D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science,* 19, 141-205.

Hintzman, D. (1988). Judgements of frequency and recognition memory in a multiple-trace model. *Psychological Review, 95*, 528-551.

Hofstadter, D. and the Fluid Analogies Research Group (1995). *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thoughts*. New York: Basic Books.

Hummel, J. & Holyoak, K. (1997). Distributed representation of structure: A theory of analogical access and mapping. *Psychological Review, 104*, 427-466.

Kokinov, B. (1988). Associative memory-based reasoning: How to represent and retrieve cases. In T. O'Shea and V. Sgurev (Eds.), *Artificial intelligence III: Methodology, systems, applications*. Amsterdam: Elsevier.

Kokinov, B. (1994a). A hybrid model of reasoning by analogy. In K. Holyoak & J. Barnden (Eds.), *Advances in connectionist and neural computation theory: Vol. 2. Analogical connections* (pp. 247-318). Norwood, NJ: Ablex

Kokinov, B. (1994b). The DUAL cognitive architecture: A hybrid multi-agent approach. *Proceedings of the Eleventh European Conference of Artificial Intelligence*. London: John Wiley & Sons, Ltd.

Kokinov, B. (1994c). The context-sensitive cognitive architecture DUAL. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kokinov, B. (1998). Analogy is like cognition: Dynamic, emergent, and context-sensitive. In K. Holyoak, D. Gentner, & B. Kokinov (Eds.), *Advances in analogy research: Integration of theory and data from the cognitive, computational, and neural sciences*. Sofia, Bulgaria: NBU Press.

Kokinov, B. & Petrov, A. (2000). Dynamic Extension of Episode Representation in Analogy-Making in AMBR. In: *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Kokinov, B. & Petrov, A. (2001). Integration of Memory and Reasoning in Analogy-Making: The AMBR Model. In: Gentner, D., Holyoak, K., Kokinov, B. (eds.) *The Analogical Mind: Perspectives from Cognitive Science*, Cambridge, MA: MIT Press.

Kolodner, J. (1984). *Retrieval and organizational strategies in conceptual memory*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Loftus, E. (1977). Shifting human color memory. *Memory and Cognition, 5*, 696-699.

Loftus, E. (1979). *Eyewitness testimony*. Cambridge, MA: Harvard University Press

Loftus, E., Feldman, J., & Dashiell, R. (1995). The reality of illusory memories. In D. Schacter (ed.), *Memory distortions: How minds, brains, and societies reconstruct the past*. Cambridge, MA: Harvard University Press.

Loftus, E., Miller, D., Burns, H. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 19-31.

Loftus, E. & Palmer, J. (1974). Reconstruction of automibile destruction: An example of interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13*, 585-589

McClelland, J. (1995). Constructive memory and memory distortions: A parallel distributed processing approach. In D. Schacter (Ed.), *Memory distortions: How minds, brains, and societies reconstruct the past*. Cambridge, MA: Harvard University Press.

Metcalfe, J. (1990). Composite holographic associative recall model (CHARM) and blended memories in eyewitness testimony. *Journal of Experimental Psychology: General, 119*, 145-160.

Minsky, M. (1986). *The society of mind*. New York: Simon and Schuster.

Moscovitch, M. (1995). Confabulation. In D. Schacter (Ed.), *Memory distortions: How minds, brains, and societies reconstruct the past*. Cambridge, MA: Harvard University Press.

Neisser, U. (1998). Stories, selves, and schemata: A review of ecological findings. In M. Conway, S. Gathercole, & C. Cornoldi (Eds.), *Theories of memory* (Vol. 2). Hove, UK: Psychology Press.

Nystrom, L. & McClelland, J. (1992). Trace synthesis in cued recall. *Journal of Memory and Language, 31*, 591-614

Petrov, A. & Kokinov, B. (1998). Mapping and access in analogy-making: Independent or interactive? A simulation experiment with AMBR. In K. Holyoak, D. Gentner, & B. Kokinov (Eds.), *Advances in analogy research: Integration of theory and data from the cognitive, computational, and neural sciences*. (pp. 124-134) Sofia, Bulgaria: NBU Press.

Petrov, A., & Kokinov, B. (1999). Processing Symbols at Variable Speed in DUAL: Connectionist Activation as Power Supply. In: Dean, T. (ed.) Proceedings of the 16th International Joint Conference on Artificial Intelligence. (pp. 846-851) San Francisco, CA: Morgan Kaufman.

Reinitz, M., Lammers, W., & Cochran, B. (1992). Memory-conjunction errors: miscombination of stored stimulus features can produce illusions of memory. *Memory & Cognition, 20*, 1-11.

Schacter, D. (1995b). Memory distortion: History and current state. In D. Schacter (Ed.), *Memory distortions: How minds, brains, and societies reconstruct the past*. Cambridge, MA: Harvard University Press.

Thagard, P., Holyoak, K., Nelson, G., & Gochfeld, D. (1990). Analog retrieval by constraint satisfaction. *Artificial Intelligence, 46*, 259-310.

Wharton, C., Holyoak, K., & Lange, T. (1996). Remote analogical reminding. *Memory and Cognition, 24* (5), 629-643.

Dissecting Common Ground: Examining an Instance of Reference Repair

Timothy Koschmann
(tkoschmann@siumed.edu)
Medical Education, 801 N. Rutledge
Springfield, IL 62711 USA

Curtis LeBaron
(lebaron@stripe.colorado.edu)
Communication Department, Campus Box 270
Boulder, CO USA

Charles Goodwin
(cgoodwin@humnet.ucla.edu)
Applied Linguistics, 3300 Rolfe Hall
Los Angeles, CA 90095 USA

Paul Feltovich (pfeltovich@siumed.edu)
Medical Education, 801 N. Rutledge
Springfield, IL 62711 USA

Abstract

How participants to a joint activity come to develop a shared or mutual understanding of what they are perceiving has long been a problematic issue for philosophers, sociologists, and linguists. We examine the abstract model proposed by Clark and Marshall (1981) whereby speakers and hearers construct mutual knowledge and by which discrepancies in definite reference are repaired. We focus in particular on forms of demonstrative reference that depend upon physical co-presence. We examine an attested example of reference repair in the operating room of a teaching hospital. It involves learning to recognize pertinent structures within endoscopic surgeries, that is surgeries in which internal spaces are rendered visible by inserting a fiber-optic lens into the body of the patient. Clark and Marshall provide a useful vocabulary for discussing referential practices in this applied setting. We are left with some questions about how to interpret certain features of their model, however. We conclude that further theoretical framing is required before we develop a full appreciation of how reference and reference repair is accomplished in day-to-day interaction.

How participants to a joint activity come to develop a shared or mutual understanding of what they are perceiving has long been a problematic issue for philosophers, sociologists, and linguists (cf., Heritage, 19884; Lewis, 1969; Sperber & Wilson, 1986; Stalnaker, 1978). One means of building "common ground" (Clark, 1996), of course, is through demonstrative reference. Even here, however, potential problems abound. When one issues the utterance "It's right here," how is it that one assures oneself that what is presented as *here* is the same as what is taken as *here* by the listener? Further, how do we detect when discrepancies have arisen and how are these discrepancies to be reconciled? We begin this paper by reviewing the pragmatic model of reference repair proposed by Clark and Marshall (1981). We then examine an instance of reference repair in an applied setting to evaluate the usefulness of this model in understanding actual referential practice.

Clark and Marshall's Model of Reference Repair

Clark and Marshall (1981) proposed an abstract model for the repair of direct references based on their proposal for how mutual knowledge is constructed. This proposal can be expressed succinctly by the following formula:

Evidence + Assumptions + Induction Schema = Mutual Knowledge[1]

where *evidence* is the grounds for the speaker and hearer's belief that both understand some matter in the same way, *assumptions* are the things taken for granted when accepting these grounds as warrants, and *induction schema* is a recursive formulation of Lewis' (1969) iterative definition of common knowledge. By this formula, evidence and assumptions are interrelated in that weaker bases of mutuality must be compensated by increasing levels of assumption. Clark and Marshall's taxonomy of evidence is broken into three categories: community membership, physical co-presence, and linguistic co-presence.[2] These evidence types, along with their associated assumptions are listed in Table I.

Mutual understanding proceeds on the assumption that speakers and listeners are each members of many different cultural communities (e.g.,

[1] In later writing (see Clark, 1996), *mutual knowledge* was expanded to *common ground*, a broader notion that subsumed mutual belief, mutual knowledge, mutual assumptions, and mutual awareness.

[2] Clark and Marshall (1981) listed indirect co-presence as a fourth category of evidence. For ease of presentation, we have condensed the categories into three.

Table 1: Bases of Common Ground
(adapted from Clark & Marshall, 1981)

Evidence	Associated Assumptions
1. Community membership	co-membership, universality of knowledge
2. Physical co-presence	
a. Immediate	simutaneity, attention, rationality
b. Potential	assumptions of 2a. + locatability
c. Prior	assumptions of 2a. + recallibility
d. Indirect potential	assumptions of 2b. + associativity
e. Indirect prior	assumptions of 2c. + associativity
3. Linguistic co-presence	
a. Potential	assumptions of 2b. + understandability
b. Prior	assumptions of 2c. + understandability
c. Indirect potential	assumptions of 3a. + associativity
d. Indirect prior	assumptions of 3b. + associativity

African Americans, soccer fans, Presbyterians, pipe fitters, speakers of French) and that membership in these communities imparts special forms of shared vocabulary and knowledge. Reference based purely on community membership assumes that the speaker and hearer hold one or more of these cultural communities in common (i.e., co-membership) and that the object of reference is known to all members of these shared communities (i.e., universality of knowledge). Clark and Marshall theorized that mutual knowledge based on community membership has an extended scope and can be carried from one conversation to another.

A second form of evidence is based on physical co-presence. When speaker and hearer are aware of an object present to both at the moment of reference (sometimes referred to as "triple co-presence"), the situation is labeled *immediate co-presence*. Although this is the strongest form of co-presence for Clark and Marshall, it too has certain assumptions. The speaker assumes that the listener is not only oriented to the object, but is also attending to it (attention) and that both are attending to it at the same time (similtaneity). It also assumes that the listener possesses the faculties to appreciate the meaning of the utterance (rationality). If only the speaker is focusing on the object, but it is available to the hearer (i.e., locatability), *potential physical co-presence* is established. If the hearer does not happen to be attending to the object of reference, but is known to have attended to it previously and can be counted upon to remember it (recallibility), then *prior physical co-presence* can be established. Attributes of components of physically co-present objects can be referred to indirectly provided the hearer recognizes (via community co-membership) the semantic links connecting the attribute or component of the object to the object (assumption of associativity).

The third category of co-presence is *linguistic*. It allows for reference to objects that have been previously introduced into the conversation. Such forms of co-presence are only prior or potential, depending on whether the object is introduced earlier or later in the stream of talk. Both types depend upon a form of assumption Clark and Marshall refer to as "understandibility." As with physical co-presence, more complex forms of linguistic co-presence are possible through association. Unlike community co-membership which is sustained over long periods, Clark and Marshall considered physical and linguistic co-presence to have relatively brief temporal extent.

As evidence for their model, Clark and Marshall direct attention to the way that speakers repair definite references. They described two forms of reference repair: *horizontal* and *vertical*. Horizontal repairs involve enhancing reference by providing additional information without altering the set of underlying assumptions. Vertical repair, on the other hand, involves advancing to a level of co-presence with fewer assumptions. For example, moving from an indirect form of co-presence to a direct form or moving from potential to immediate co-presence or shifting from linguistic to physical co-presence. Because community co-membership has assumptions that are entirely different from those underlying physical and linguistic co-presence, it allows only for horizontal forms of repair.

The model of reference repair presented by Clark and Marshall was largely linguistic. Clark (1996) later elaborated on the notion of common ground. He made a conceptual distinction between *communal common ground*, something that rests largely on community co-membership, and *personal common ground*, with a correspondence to what has been previously described as physical and linguistic co-presence. He expanded his treatment of personal common ground to include "joint perceptual experiences" and "joint actions" (p. 112), that is gesticulation, observed actions, and other

features of the social setting in addition to talk. As we turn to an instance of actual reference repair in an applied setting, we see the importance of taking a broader and more situated view of referential practice. In particular, we begin to see some of the complexities embedded in certain features of Clark and Marshall's model, such as the assumption of locatability.

Analyzing Reference in an Endoscopic Surgery

The setting within which we have chosen to study referential practice is the operating room (OR) of a busy teaching hospital. Within this context, there are multiple forms of work being performed simultaneously. On the one hand, there is a cycle of activity surrounding the performance of a particular surgical procedure itself within which each of the members of surgical team plays a specific role. At the same time, there is instructional work to be done as well. In the fragment to be analyzed here, one participant ("Attending") is a highly-experienced surgeon, ultimately responsible for the safe and successful outcome of the surgery. A second ("Resident") is a surgeon in the final year of his surgical residency, who had by his own estimate participated in 80 to 90 surgeries of the type to be described here (by comparison, the attending surgeon reported that he has performed 1200-1300 of these surgeries over the course of his career). The remaining participant ("Clerk") is a third-year medical student enrolled in a clerkship rotation. This was his first surgical experience. Attending, therefore, is providing guidance and supervision to the resident and both Attending and Resident are responsible for providing instruction to the medical student.

The surgical procedure in which they are engaged is a laparoscopic cholecystectomy, that is the removal of the gall bladder with the aid of an endoscopic camera. Such surgeries were of interest to us because of the manifold challenges to perception and coordination that they pose to participants. Surgeons are called upon to translate what they see on a 2-D TV monitor into a model of what is happening within the not directly inspectable belly of the patient. The image seen on the screen is a magnified view that facilitates precise manipulation on the part of the surgeon, but can be disorienting for newcomers. The orientation of the view on the screen is arbitrary, though the convention is to orient the lens in such a way that the projected image most closely resembles what would be seen in an open surgery (that is a ventral view in which up is anterior and down is posterior). Since participants on opposite sides of the operating table observe different monitors, however, the person assisting the surgeon from the opposite side of the table receives an inverted view.

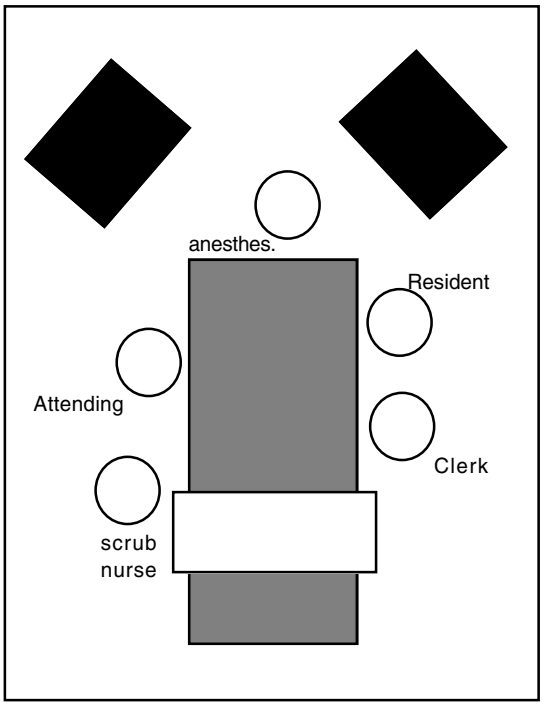


Figure 1. Arrangement of the operating room.

Looking within the endoscopic space is a team effort with different members responsible for operating the camera, "retracting" obstructing organs, and conducting the surgery. This requires substantial coordination in that a view of the workspace adequate to carry out the procedure can only be achieved if all members of the team correctly anticipate the needs of the surgeon. Although the participants work in close proximity to one another, many of the normal resources for effecting mutual orientation are not available to them. Their hands, for example, are occupied much of the time and, as a consequence, cannot be employed for gesture. Further, because they do not attend to the same monitors and because the monitors are located at a distance from where they work, it is difficult for them to use each others' gaze as a cue for orientation as is often done in more typical face-to-face interaction (Goodwin, 1986; Streeck, 1993, 1994).

Attending assists the resident from the left side of the table (see Figure 1). Clerk, standing to the left of Resident on the right side of the table, controls the rod lens of the endoscopic camera. The surgery is considered routine. It consists of isolating the small duct (the cystic duct) through which the gallbladder empties into the common bile duct and the vessel (the cystic artery) that supplies the gallbladder with blood, ligating both with surgical clips, and severing them. The gall bladder is then gently teased from the liver and extracted through one of the "ports" in the abdominal

wall. The greatest technical challenge is correctly identifying the cystic duct and cystic artery, as serious post-surgical complications may arise were clips to be applied to the wrong structures.

Repairing Reference in the OR

Space restrictions prevent us from presenting here a full analysis of the interaction. A more detailed analysis of the fragment can be found elsewhere (Koschmann, Goodwin, LeBaron, & Feltovich, in prep). A transcript can be found in Appendix A. [3] It begins (lines 1-9) with Attending describing the surgical procedure to Clerk. At the same time and throughout the course of this interaction, the resident was performing a blunt dissection to expose the cystic duct and the cystic artery using the tool in his right hand (a "black grasper"). This dissection was performed by burrowing the tip of the grasper into a bundle of connective tissue binding the bottom edge of the gall bladder to the common bile duct and then gently spreading apart the jaws of the instrument. Attending and Clerk observed his progress on their respective monitors.

The expression *cystic artery* is introduced here for the first time (line 2). In terms of Clark and Marshall's model, Attending's use of this expression is authorized by Clerk and Attending's membership in some common community. Attending displays by his choice of language a set of presuppositions about what would be understandable to a third-year medical student. Resident's first demonstrative reference to the cystic artery (line 5) specifies a region in which the cystic artery can be found, though it may not necessarily be visible at the moment in which he makes the reference. In Clark and Marshall's terminology, therefore, these references signal potential physical co-presence. This raises interesting questions about what the assumption of locatability means in this particular situation, however. If it means that the cystic artery is simply available to Clerk's viewing, Resident's utterance would suggest that he believed the cystic artery to be locatable at the moment of reference. If one has never seen a cystic artery on an endoscopic display, however, is it still locatable there?

Clerk's query in line 10 makes visible his orientation to unfolding process. The cystic artery may or may not be visible at that point in time, but his use of the adverb *yet* expresses a confidence that it will eventually be made manifest to all. Attending's reply in line 13 ratifies this view. Like Goodwin's (1999) archeologists excavating through sedimented strata of soil, surgeons must dissect through various layers of anatomical structure. They speak of *planes of dissection*, meaning the surfaces available to sight at specific junctures within a procedure. In an endoscopic surgery, however, the cystic artery will never be physically co-present in the same way that it would in an open surgery since its presence is mediated through a video viewing system. Attending's deictic particle *here*, therefore, anchors not to the conventional origo of the speaker's corporal location, but rather to a virtual origo located in the shared media space.

Resident eventually provides six separate demonstrations of the cystic artery before receiving a tentative sign of recognition on the part of Clerk (line 19).[4] Learning to locate pertinent structures on the video display is an important aspect of "professional vision" (Goodwin, 1994). Resident's *there* (line 18) was coordinated with a point to a white stripe within the bundle of connective tissue being viewed. Although gesture is often characterized by linguists as supplementing speech, Hindmarsh and Heath (2000) described instances in which "The deictic term segments the gesture, displaying just the moment at which it is sequentially relevant" such that "the talk reflexively works on behalf of the gesture" (p. 15).

Resident's repeated efforts to demonstrate the cystic artery, could be described in terms of Clark and Marshall's model of reference repair as an attempt to eliminate the assumption of locatability. That is, he was striving to promote his shared knowledge with Clerk from potential to immediate co-presence. But what does it really mean to be "locatable." The whole idea of "professional vision" is to acquire the ability to see as presumably more-skilled others can see. If locatability assumes not only that the listener can see (in the sense of having adequate vision, an unblocked view, etc.) what is visible to the speaker, but must also be able to see in the same ways as the speaker (i.e., share the speaker's "professional vision"), then it becomes a very complex kind of assumption, in many ways just as complex as the thing it sets out to explain, namely mutual understanding.

As the fragment continues, Attending raises some concerns about Resident's identification of the cystic artery. On paper, Attending's "That may be right" (line 27) might be construed as a tentative positive appraisal. Resident's reply (line 29), however, treats it as an incomplete utterance, as in "That may be right [hepatic]." Resident's efforts to achieve mutual understanding with Clerk, therefore, have revealed a potential discrepancy in understanding among Resident and Attending. The fragment concludes with Resident

[3] The transcription conventions used here are described in Atkinson and Heritage (1984).

[4] Resident's demonstrative reference in line 3 ("Right there") is heard to be referring to the cystic duct, a topic of discussion prior to the transcribed segment, rather than the cystic artery.

and Attending resolving to search further for the cystic artery.

Discussion

Here in a nutshell we see the problem of mutual knowledge. Resident takes some pains to demonstrate to Clerk what he (Resident) believes to be the cystic artery. After some prompting, Clerk declares that he now sees it. Other than his avowal, however, we have no evidence that he indeed sees what Resident has taken such trouble to display. In demonstrating for Clerk what he has taken to be the cystic artery, however, Resident has inadvertently made visible a discrepancy in his presumed common ground with Attending (or, at the very least, a difference in their levels of confidence that the indicated structure is in fact the cystic artery). Clark (1996) defined *grounding* as establishing a claim "as a part of common ground well enough for current purposes" (p. 221). For the purposes of Clerk's instruction, the exchange would seem to have provided ample grounding for his understanding. However, for the purposes of conducting a safe surgery, the concerns raised by Attending might suggest that more grounding is required.

Clark and Marshall provide a useful vocabulary for discussing referential practices in this applied setting. Their model of reference repair, however, hinges upon a calculus of assumption maintenance and herein lies the rub. The conceptual difficulties of mutual knowledge that their model was meant to address have not been completely dispelled, but, instead, arise in new forms when we look more carefully at the underlying assumptions. As we have seen, the assumption of locatability can be quite complex when examined *in situ*. We are in full accord with Clark's shift from a treatment of reference as a simple matter of linguistic interpretation to a more situated model that encompasses "joint actions" and "joint perceptual experiences" and we think that this will lead to a richer understanding of concepts like locatability. For one thing, it would help to illuminate how participants' own unfolding activities contribute to the determinant sense of what is seeable at any given moment. Furthermore, we have much to learn about the interactions between different kinds of bases of shared understanding. Professional vision, for example, draws upon the associated assumptions of both community membership and physical co-presence.

In a situation in which the establishment of common ground is essential, we see just how elusive shared understanding can be to achieve. Our analysis of the fragment of interaction in the OR would suggest that we have a way to go before fully appreciating how these factors enter into our day-to-day practices of reference and reference repair.

Acknowledgements

We would like to acknowledge the kind assistance of Drs. Barnett, Dunnington, Dutta, Reisman, and Thiele for providing access to this workplace. We also thank Mindy Conlee's for her transcription assistance. Finally, we thank Randi Engle and two anonymous reviewers for their useful comments.

References

Atkinson, J.M. & Heritage, J. (1984). Transcription notation. In J.M. Atkinson & J. Heritage (Eds.), *Structures of Social Action* (pp. ix—xvi). NY: Cambridge University Press.

Clark, H. (1996). *Using language*. NY: Cambridge University Press.

Clark, H. & Marshall, C. (1981). Definite reference and mutual knowledge. In A.K. Joshi, B.L. Webber, & I.A. Sag (Eds.), *Elements of discourse understanding* (pp. 10-63). NY: Cambridge University Press.

Goodwin, C. (1986). Gestures as a resource for the organization of mutual orientation. *Semiotic*a, *62*, 29-49.

Goodwin, C. (1994). Professional vision. *American Anthropologis*t, *96*, 606-633.

Goodwin, C. (1999). Pointing as situated activity. Paper presented at meeting of Society for Social Studies of Science.

Heritage, J. (1984). *Garfinkel and Ethnomethodology*. Cambridge, U.K.: Polity Press.

Hindmarsh, J. & Heath, C. (2000). Embodied reference: A study of deixis in workplace interaction. *Journal of Pragmatics*, *32*, 1855-18788.

Koschmann, T., Goodwin, C., LeBaron,, C., & Feltovich, P. (in prep). A mediated conversation on the cystic artery: Reflexivity and the 'inscrutability of reference'. *Human Computer Interaction.*

Lewis, D. (1969). *Convention*. Cambridge, MA: Harvard University Press.

Sperber, D. & Wilson, D. (1986). *Relevance*. Cambridge, MA: Harvard University Press.

Stalnaker, R.C. (1978). Assertion. In P. Cole (Ed.), *Syntax and semantics: Pragmatics, Vol 9* (pp. 315-332). NY: Academic Press.

Streek, J. (1993). Gesture as communication I: Its coordination with gaze and speech. *Communication Monographs, 60*, 275—299.

Streek, J. (1994). Gesture as communication II: The audience as author. *Research on Language and Social Interaction, 27*, 239—267.

Appendix A: "Can you see the cystic artery yet?"

Tape: #99-001b (0:08:48:10 to 0:09:55:28)

ATT: Yeah (.) the other “thing to do:: is make sure you have your cystic (.) artery out

RES: ”(Right there)

CLK: Uh huh

RES: ““Which is right back in here‘

ATT: ””That way there is nuthin ’ else before you h- (.) hit the edge of the liver

(1.6)

ATT: That (kinda) guarantees you're safe too

CLK: Can you see the cystic artery yet? I:'d=

RES: =It's §r::ight back in the:re

(2.1)

ATT: º(We'll)º get it out here in a minute.

(0.2)

RES: See it right there?

(0.2)

CLK: U::mmmmm=

RES: =Right (0.2) “there

CLK: ”Okay yeah (.) yeah

(0.2)

RES: That looks like the (0.2) where the money's a:t

CLK: Uhkay

(0.2)

RES: En yih can see it's hanging out in that “tract.

ATT: ”º(That's) actually big º

ATT: That's pretty bi:g, that may be ri:ght

(0.4)

RES: That's right hepatic? 30 (1.2)

ATT: ““(Comin' up)

RES: ””The cystic may be up a little higher?=

ATT: =(Yup)

RES: You can see how easy that is to do we were just talking about that.

CLK: So you jus' dissect until you'r:e absolutely sure

ATT: Ah “hah

RES: ”Yeah=

CLK: =Till you see both the right hepatic and the cystic and then (4.6)

RES: Remember how we talked about there's no:: (.) no collateral flow beyond (.) the (geh) the right gastric so if you happen to clip the right (.) you're kind of in trouble, (even) you've clipped an end artery.

(2.5)

ATT: I don't know.

RES: Want me to take the duct down?

RES: And look up in here (.) for the (.) cystic artery?

Kinds of Kinds: Sources of Category Coherence

Kenneth Jeffrey Kurtz (kjk@northwestern.edu)
Dedre Gentner (gentner@northwestern.edu)
Department of Psychology, Northwestern University, 2029 Sheridan Rd
Evanston, IL 60208-2710 USA

Abstract

A fundamental question in the study of concepts is what makes sets of examples cohere as categories. We present results of three studies designed to compare standard taxonomic categories with categories that take their meaning from relationships extending outside of the individual example. An exemplar generation task is used to differentiate relational categories from taxonomic kinds and to compare possible subtypes of extrinsically cohering categories based on goals or thematic contexts. Results provide strong support for the intrinsic—extrinsic distinction and reveal signatures of underlying organization among the types of categories investigated.

Introduction

Categories play a fundamental role in cognition. The internal structure of categories supports numerous functions including classification, prediction, and reasoning. Categories give rise to an extension: the set of examples in the world that are members. The coherence of a category is the meaningful basis according to which these members go together.

One traditional view holds that the correlational structure of the environment determines category coherence due to systematic patterns of within-category similarity and between-category difference (Rosch & Mervis, 1975). Murphy and Medin (1985) propose the theory-based view that challenges the idea that similarity itself explains category coherence. 'Respects' for similarity (i.e., a basis for the selection of features and weights) must be specified in order for concepts to exist as groups of *like* examples. Additionally, they argue that category representations are richer than lists of features and must include relationships that hold within and between examples of categories. The tension inherent in the need for a constrained, yet rich basis of category coherence poses a continued challenge to theorists (Goldstone, 1994).

A useful source of inspiration is structural alignment theory—which has proven successful as an account of comparison processes such as similarity and analogy (Gentner, 1983; Markman & Gentner, 1993; Gentner, Rattermann, & Forbus, 1993). This framework offers a perspective for addressing the question of category coherence in a manner in keeping with the theory view. Respects for similarity can arise from the process of aligning corresponding predicates of two structured representations (Gentner & Markman, 1997; Medin, Goldstone, & Gentner, 1993; Markman & Wisniewski, 1997). Relational similarity drives the alignment process and largely determines the quality of a match. In the same way that structural alignment theory looks to shared relations (more than attributes or objects) to explain similarity, we can look to relationships between objects as a source of category coherence. While theory-based categories may cohere around intrinsic relationships (like the causal link between genetic material and physical features), we focus here on relationships extending beyond the individual example. For instance, the category *barriers* consists of examples that conform to the relationship: BLOCKS (X, Y).

Barr and Caplan (1987) distinguish between the intrinsic properties of a category which are true of an example in isolation versus extrinsic features which hold only in relation to other objects. As an alternative to category members bound together by common intrinsic structure (relations or attributes), category coherence can be derived from relations extrinsic to individual examples. The extreme case of extrinsic coherence is relational categories like *barriers*—members cohere based on fulfilling a core relationship. The roles of X and Y in the blocking relation can be filled by anything—so as long as the relationship holds, membership is secure. The examples of a relational category may have few or no intrinsic properties in common with one another. In this sense, relational categories are akin to analogies. Both 'prison bars' and 'raging river' are members of the category *barrier*, despite their sharing no intrinsic similarity.

Another case in which the category coherence is extrinsic is Barsalou's (1983, 1985) ad-hoc or goal-derived categories which are organized around ideals (properties that optimally promote goal resolution) rather than central tendency. Again, categories such as *things to take out of the house in case of fire* violate the correlational structure of the environment since member examples have few properties in common. Goldstone (1995) makes a useful distinction between default and directed similarity; where the former is the basis of graded structure and broad inferential power of taxonomic categories, while the latter is the focal, context-specific sense of similarity underlying ad-hoc categories or analogical relationships.

Categories may also be grounded by properties beyond the individual example that are not specific relationships. As an example, consider *items associated*

with working at an office desk. Thematic categories consist of examples that tend to cluster or co-occur in particular contexts. As Wisniewski et al. (1996) noted, such categories are often expressed as mass noun superordinates – e.g., *groceries* or *workout equipment*. There may be other relations in addition to spatiotemporal contiguity between particular pairs within a thematic category, but the members need not share any particular similarity.

In the present work, we use an exemplar-generation task to investigate and compare these possible bases for what makes things "go together" as a category. The sources of category coherence are: default similarity to the central tendency, directed similarity to ideals or to a core relationship, and patterns of contiguity in spatio-temporal contexts. We believe that there are kinds of categories that are best explained in terms of each of these sources of coherence, while there are also categories that are grounded in mixed forms of coherence. Barsalou (1985) shows that ideals account for a significant portion of the variance in typicality not only for goal-derived categories, but also taxonomic categories. Furthermore, different kinds of experts have been shown to organize the taxonomic category *tree* in terms of ideals derived from their experience (Lynch, Coley, & Medin, 1999).

Our main goals are: 1) to assess the psychological reality of extrinsically cohering categories in contrast with the intrinsic coherence attributed to standard taxonomic categories; 2) compare types of extrinsically cohering categories; and 3) address the non-uniformity of coherence in real-world categories. We use an exemplar generation task along with several follow-up measures to determine whether these different posited sources of category coherence are made evident in the behavior of the category.

Experiment 1: Exemplar generation

We begin by using exemplar generation as a means of indexing category coherence. Through measuring the content and dynamics of responding, a picture can develop of the organization of the knowledge being accessed. This technique has been used sporadically in the categorization literature. Goal-derived categories have been shown to support exemplar generation, but produce less output and show a lesser degree of correlation of output dominance with typicality than taxonomic categories under short time intervals (Barsalou 1983, 1985; Vallee-Tourangeau, Anthony, & Austin, 1998). In addition, greater output consensus was found for taxonomic than ad-hoc categories.

Several lines of evidence lead to the prediction that taxonomic categories should be easier and more natural than relational categories. As noted above, relational categories are akin to analogies: their members need share only relational similarity, not overall literal similarity. In contrast, the members of taxonomic categories share overall similarity. For example, two instances of *vegetable* are likely to have considerable intrinsic similarity (seeds, skin, etc.), as well as some extrinsic similarity (sold in stores, provide nourishment, for people, etc.). There is considerable evidence that relational similarity is more difficult to access in memory than object similarity (Gentner et al, 1993; Holyoak & Koh, 1995); and acquired later in development (Gentner & Rattermann, 1991). Further, Barr & Caplan (1987) showed that categories characterized by extrinsic features possess a greater degree of graded structure—possibly due to lower category validity of extrinsic properties. Thus we expect that relational categories will be less fluent (fewer runs of responses with minimal inter-item delay), less generative (fewer responses produced), and less consistent (lower agreement between participants) than taxonomic categories.

Evidence on the behavior of goal-derived categories leads us to expect that they should also be less generative than taxonomic categories. The investigation of thematic categories is more exploratory. The fact that members of thematic categories – such as 'ticket' and 'popcorn' for *things associated with going to the movies* -- lack not only intrinsic, but even relational similarity, might lead one to expect low generativity. On the other hand, the fact that thematic associates share spatiotemporal contiguity suggests that members might readily prime one another.

Method

Participants. 75 undergraduates from Northwestern University served as participants in order to fulfill an introductory course requirement.

Materials and design. Eight category cues (see Table 1) were selected for four different types of categories: taxonomic (all count noun superordinates), thematic, goal-derived, and relational. Natural language labels for the categories were determined for optimal clarity. Each participant generated exemplars for two category cues of each type. Item assignment was accomplished by random selection. The experiment used a within-Ss design with four item conditions corresponding to the types of categories.

Procedure. Participants read a set of instructions appearing on the computer screen. They were told they would be shown category cues (words or phrases) and asked to generate as many examples as they could during 4-minute intervals. To illustrate the nature of the task, a sample was shown: examples of *beverage* include water and milk. Participants were asked to work

as quickly and accurately as possible. The category cue remained visible for the duration of the trial. Participants typed into a response window on the screen. For each response, the time of the initial keystroke and the time of typing the return key were recorded. All responses entered for that cue remained accessible to the participant in a history window. This is comparable to a pencil-and-paper version of a listing task, but allowed recording of precise timing information. The eight category cues were presented in a random order. The entire experiment took approximately 35 minutes.

Table 1: Categories used.

Taxonomic	an animal
	a plant
	a fruit
	a vegetable
	a vehicle
	a household appliance
	a type of dwelling
	a musical instrument
Thematic	an item associated with:
	dining out at a restaurant
	going to the movies
	working at an office desk
	preparing for sleep at night
	working out at the gym
	going to the beach
	a party
	taking an airplane trip
Goal-derived	an item to take on a camping trip
	an item to remove from the house in case of fire
	an item not to eat while dieting
	a picnic activity
	a thing to do for weekend entertainment
	a way to advertise something
	an item to sell at a garage sale
	a thing that makes someplace desirable to live
Relational	a weapon
	a trap
	a guide
	a signal
	a barrier
	a tool
	a filter
	a shield

Results

The results are summarized in Table 2. All analyses were conducted by item since each participant only responded to two out of the eight items in each condition. The number of presentations of each item was not equal due to the random selection, but the number of presentations of each type of item was equal. All items were presented at least seven times.

Two analyses of response dynamics were performed on the entire data set (N = 75): response fluency and clustering. *Response fluency* was a measure of how long it took to generate each response. The "downtime" between any two responses was computed as the amount of time between the initial keystroke of each response. Item fluency was determined by the median downtime between responses. Mean fluency (in milliseconds) varied across item condition, as shown in Table 2. Comparison of the means using a one-way ANOVA showed a reliable difference between conditions, $F(3,31) = 6.44$, $p = .002$. As predicted, the Relational condition ($M = 10832$) was significantly less fluent according to post-hoc comparisons (all such tests we report were performed using the Bonferroni correction) than both the Taxonomic ($M = 6370$), $p < .01$) and the Thematic ($M = 6793$), $p < .01$) conditions.

We assessed *clustering* of responses in several ways yielding convergent results. In one analysis, any two responses that occurred within less than 67% of the median downtime for all responses by that participant were considered to be clustered together. To measure the degree of clustering, a ratio was constructed between the number of clustered responses and the number of isolated responses. A value greater than one indicates more clustered than isolated responses.

The clustering ratios are shown in Table 2. A one-way ANOVA was used to assess the differences between conditions, $F(3,31) = 6.76$, $p = .001$. Post-hoc comparisons showed reliably less clustering for Relational ($M = .64$) than for Taxonomic categories ($M = 1.38$), $p < .02$. In addition, significant differences were found between Relational and Thematic ($M = 1.48$), $p < .01$, as well as between Goal-derived ($M = .82$) and Thematic ($p < .04$).

Table 2: Summary of Results of Experiment 1.

	Tax	Thematic	Goal	Relation
Productivity	23.2	21.9	19.1	14.2
Consensus	25.8	18.1	17.3	13.8
Paragons	5.1	2.1	1.1	0.8
Easy-Access	2.5	3.0	1.5	0.6
Fluency	6370	6793	8140	10832
Clustering	1.4	1.5	0.8	0.6

Analyses of response content. A subset of the responses (the first 46 of the 75 participants) was analyzed intensively using a scoring procedure performed by trained undergraduate research assistants. Responses were removed from the analysis on the basis of a clear failure to understand the task or to undertake it seriously. Repeated and blank responses were also removed. A conservative coding of responses was performed: pure synonyms, abbreviations, and minor

syntactic variations (e.g., singular versus plural) were treated as the same response.

Productivity or item output was measured as the mean number of responses produced. A one-way ANOVA was performed to test for differences between Goal-derived (M = 19.1), Relational (M = 14.2), Taxonomic (M = 23.2), and Thematic (M = 21.9). An effect of item condition on productivity was found $F(3,31) = 3.87$, $p = .02$. The effect appears to be driven by the low mean productivity in the Relational condition. Post-hoc comparisons revealed a significant difference between Relational and Taxonomic ($p < .03$). A marginal difference was also found between Relational and Thematic ($p < .07$).

In order to evaluate whether participants tended to generate the same responses to the categories, output *consensus* was measured in two ways. For each item, the percentage of participants who produced each response was computed and the mean was taken across all responses generated for that item. By this measure, output consensus varied as follows: Goal-derived (M = 17%), Relational (M = 14%), Taxonomic (M = 26%), and Thematic (M = 22%). A one-way ANOVA showed a reliable difference between item conditions $F(3,31) = 6.41$, $p = .002$. Post-hoc comparisons showed significant differences between Taxonomic and both Relational ($p = .001$) and Goal-derived ($p < .03$). The difference between Taxonomic and Thematic was marginally significant ($p < .07$).

As a convergent measure, output consensus was also analyzed by computing the percentage of responses that occurred frequently (generated by at least 60% of the participants receiving the item). Few responses were widely agreed upon by participants: Goal-derived (M = 4%), Relational (M = 2%), Taxonomic (M = 12%), and Thematic (M = 5%). Group means were compared using a one-way ANOVA that revealed an effect of condition on output consensus $F(3,31) = 4.50$, $p = .01$. On both measures, agreement was greatest for Taxonomic and lowest for Relational.

The content of the exemplar generation data showed particular responses that occurred with great regularity (produced by at least 85% of participants). To give an example from each type of category: 'car' was a paragon of the Taxonomic category *vehicle*, 'wall' was a paragon of the Relational category *barrier*, 'tent' and 'sleeping bag' were paragons of the Goal-derived category *an item to take on a camping trip*, and 'check' was a paragon of the Thematic category *an item associated with dining out at a restaurant*. The prevalence of such high-agreement responses was computed by counting the number of paragons for each item. This measure is reported as a frequency, not as a percentage of the total set of responses, since the presence of special responses is not likely to follow from the overall breadth of responding. The mean number of paragons is shown in Table 2. A one-way ANOVA showed a significant difference between groups, $F(3,31) = 4.36$, $p = .01$. As confirmed by post-hoc comparisons, Taxonomic categories yielded significantly more paragons than Relational ($p < .02$) or Goal-derived ($p < .04$).

In addition, certain responses were found to occur both early and often in the exemplar generation task. The presence of such easy-access items was determined according to mean list position (normalized by list length). Frequent responses with a mean position score of less than 0.3 were considered easy-access responses. Easy-access responses sometimes, but not always, corresponded with category paragons. For example, the easy-access responses for the Taxonomic category *vehicle* included the paragon 'car' plus 'truck.' 'Wall' was the only paragon as well as the only easy-access response for the Relational category *barrier*. 'Tent' and 'sleeping bag' were both paragons and easy-access responses for the Goal-derived category an *item to take on a camping trip*. For the Thematic category *an item associated with dining out at a restaurant,* the paragon was 'check', but the easy-access responses were 'waiter' and 'menu'. A one-way ANOVA showed an effect of item condition on the frequency of easy-access responses, $F(3,31) = 5.98$, $p < .005$. Post-hoc comparisons showed Relational categories produced reliably fewer easy-access items than Thematic ($p < .005$) and Taxonomic ($p < .03$) categories.

Discussion

A basic pattern can be discerned across the set of results. The Relational and Taxonomic categories are reliably different on nearly every measure tested. This provides strong support for the predicted differentiation of intrinsic and extrinsic forms of category coherence. The analysis of response content reveals that Relational categories (and to a lesser degree Goal-derived categories) are less productive, less consistent, and less likely to have paragons or easy-access responses. The thematic categories are distinct in the high frequency of easy-access responses.

Experiment 2a: Pairwise Similarity of generated exemplars

We suggested above that only the Taxonomic categories possess coherence based on intrinsic similarity. To confirm this claim in terms of the actual responses generated by participants, we obtained similarity ratings for within-category pairs.

Method

Participants. 37 undergraduates from Northwestern University served as participants in order to fulfill an introductory course requirement.

Materials. Stimulus materials were sets of the six responses generated with the highest consensus on half of the category items in Experiment 1 (the half selected were those items yielding the most high-consensus responses). All within-category pairs were tested.

Procedure. Participants received instructions to rate the similarity of pairs of items on a scale from low (1) to high (5). An example of both high and low similarity was provided. All possible within-category pairs were presented in pseudo-random order (no pairs from the same category were presented consecutively). Each pair was presented in random left-right order. Participants used the mouse to click on the button labeled with the numerical rating. A response could be changed by re-selecting before clicking "OK" to continue.

Results and Discussion

Mean pairwise similarity was computed across the fifteen within-category response pairs for each item. Across all participants, Taxonomic (M = 3.8) pairs showed the highest mean similarity while the other conditions were nearly equal: Relational (M = 2.3), Goal-derived (M = 2.4), Thematic (M = 2.4). This difference was confirmed by a one-way ANOVA, $F(3,15) = 13.31$, $p < .001$. Post-hoc comparisons showed highly significant differences between Taxonomic and the other conditions (all $p < .01$).

The results are consistent with our prior findings in showing an advantage for Taxonomic categories over Relational categories. In addition, neither Thematic nor Goal-derived categories showed high within-category similarity. This pattern is consistent with the view that taxonomic categories are based on overall default similarity while the other types are grounded in alternate forms of category coherence.

Experiment 2b: Category transparency of generated exemplars

A measure of the nature of a category's coherence is how readily the common basis can be perceived given a large set of examples. In this study, participants were presented with sets of generated category examples and asked to say what they had in common. As before, an advantage was predicted for Taxonomic categories.

Method

Participants. 25 undergraduates from Northwestern University served as participants in order to fulfill an introductory course requirement.

Materials. The same materials were used as in Experiment 2a. Instead of pairs, the six high-consensus responses for each category were presented together. A packet was prepared with one page for each category. The six responses were displayed on the page in three staggered columns of two to minimize spatially organized sub-groups within the set. The order of the six responses was fixed (alphabetical). The pages of each packet were randomly ordered.

Procedure. On each page of the packet to be completed, participants were given a blank line on which to answer: "What would you say the following examples have in common?" Additionally, three blank lines were provided to: "Try to think of a few more examples that fit well with the group."

Results and Discussion

The results of the commonality task are of principle interest since participants were almost always able to generate consistent additional examples. Participants routinely interpreted the commonality judgement as if the task were to induce the category from the examples. Category transparency for each item was computed as the percentage of participants whose response was scored as a match to the initial category cue used in Experiment 1. Responses that captured the meaning of the category, but differed in word choice, were accepted as matches. Taxonomic (M = 100%) and Thematic (M = 97%) items showed very high mean category transparency. Relational (M = 74%) and Goal-derived (M = 68%) items showed considerably less transparency. A one-way ANOVA revealed an effect of item condition, $F(3,15) = 3.71$, $p < .05$. A planned contrast between Relational and Taxonomic showed a reliable condition difference, $t(12) = 2.19$, $p < .05$.

Taxonomic categories are highly transparent, as should follow from their high intrinsic similarity. In contrast, Relational categories, as expected from their extrinsic similarity grounding, show lower transparency—likewise for Goal-derived categories. While participants were sometimes able to instantiate these original categories from the bottom-up, there were frequent failures as well. The above three types behaved consistently on pairwise similarity and category transparency tasks. However, Thematic categories showed a marked difference: despite low inter-item similarity (Experiment 2a), the connection among the group as a whole was highly transparent. We conjecture that the multiple examples in the current task invited participants to instantiate unifying spatiotemporal contexts.

General Discussion

Results across the two studies strongly bear out our predictions. Taxonomic categories show high intrinsic similarity and all the many advantages in terms of fluency and generativity which follow. Relational categories are markedly less similar, less transparent,

and less generative. The remaining two kinds of categories are intermediate. Goal-derived categories often pattern with Relational categories—not surprisingly, since relations link objects to goals. Thematic categories are in some sense the outlier; while they are highly fluent, they are grounded not in commonality, but in associativity.

Before discussing the implications of these data, there are some concerns to be addressed. Our choice of items represents our best effort to capture each type, but some factors were not precisely controlled. One issue is whether lower production can be attributed to smaller set size. Unfortunately, establishing the size of a relational category is not straightforward; e.g., should examples such as 'lack of education' (listed under barrier) be included? Relational categories may include more abstract or less familiar examples. These factors could play a role in generation.

To summarize, traditional explanations of real-world categories have appealed to feature overlap and the correlational structure of the environment. The emphasis suggested by the theory view of concepts on relations within and between category examples and the success of the structural alignment account of psychological similarity point toward a key role for relations underlying category coherence. The research reported here shows that extrinsic coherence can support categorical organization and points to individual signatures for different kinds of categories.

Given that relational categories appear to bring up the rear on all our measures, should we draw the implication that such categories are not psychologically real or natural? We would answer No, and Yes. Relational categories are indeed less natural than categories based on overall similarity; they do not provide a first-order basis for making sense of the world. But they provide structural organizers for understanding the world in ways that cross-cut object-based categories. We suggest that categories such as *barrier*, *operator*, and *catalyst*, though they may never be as facile as object categories, pay their way as tools of cognition.

Acknowledgments

This research was supported by an NIH-NRSA post-doctoral fellowship to Ken Kurtz and by NSF Grant SBR-95-11757 to D. Gentner. This paper was partially prepared while D. Gentner was a fellow at the Center for Advanced Study in the Behavioral Sciences with support from the William T. Grant Foundation, award #95167795. We thank Melissa Wu for contributions to the design of experiment, Adrienne Rosen, Evan Ransom, and Jessica Goethals for help in processing the data, Jeremy Cloud for software design, and Kathleen Braun for her help throughout the project, and the Similarity and Analogy group at Northwestern.

References

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition, 11,* 211-227.

Barsalou, L.W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 11,* 629-649.

Barr, R.A, & Caplan, L.J. (1987). Category representations and their implications for category structure. *Memory & Cognition, 15,* 397-418.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7,* 155-170.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist, 52,* 45-56.

Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition, 65,* 263-297.

Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability and inferential soundness. *Cognitive Psychology, 25*, 524-575.

Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition 52,* 125-157.

Goldstone, R.L. (1995). Mainstream and avant-garde similarity, *Psychologica Belgica, 35,* 145-165.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition, 15,* 332-340.

Lynch, E. B., Coley, J. D., and Medin, D. L. (1999). Tall is typical: Central tendency, ideal dimensions, and graded category structure among tree experts and novices. *Memory & Cognition, 28,* 41-50.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology, 25,* 431-467.

Markman, A. B., & Wisniewski, E. (1997). Similar and different: The differentiation of basic-level categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 54-70.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100*, 254-278.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289-316.

Rosch, E., & Mervis, C.B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7,* 573-605.

Wisniewski, E. J., Imai, M., & Casey, L. (1996). On the equivalence of superordinate concepts. *Cognition, 60,* 269-298.

Vallee-Tourangeau, F., Anthony, S.H., Austin, N.G. (1998). Strategies for generating multiple instances of common and ad hoc categories. *Memory, 6,* 555-592.

Learning Perceptual Chunks for Problem Decomposition

Peter C. R. Lane (PCL@Psychology.Nottingham.Ac.Uk)
Peter C-H. Cheng (Peter.Cheng@Nottingham.Ac.Uk)
Fernand Gobet (FRG@Psychology.Nottingham.Ac.Uk)
ESRC Centre for Research in Development, Instruction and Training, School of Psychology,
University of Nottingham, University Park, Nottingham NG7 2RD, UK

Abstract

How students learn to use diagrammatic representations is an important topic in the design of effective representations for problem solving or conceptual learning, but few good models of their learning exist. In this paper, we explore the learning process with an experiment using AVOW diagrams as a representation for solving problems in electric circuits. We find that the participants decompose each circuit into a similar set of groups when solving the problems. The natural question is whether these groups are an artifact of the visual form of the circuit, or indeed the result of prior learning. We argue that the decompositions are a result of perceptual chunking, and that they are (at least partly) a result of learning. In support of this, we describe a computational model of perceptual learning, CHREST+, and show that it predicts the decomposition of problems evident in the participants' data.

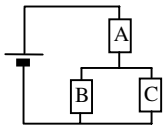
Introduction

The role of effective representations in supporting or enhancing the conceptual understanding of a student is an important topic within educational psychology (Cheng, 1999b; Larkin & Simon, 1987). However, in spite of the educational interest in effective representations, the manner in which students learn with and about different representations is not well understood. The traditional method of looking for chunks, through timing information (e.g. Chase & Simon, 1973), is hard to apply in problem-solving tasks, as the timing information is associated with the solution, and not directly linked to the problem. We instead use a computational model to match the solutions produced by students in a typical diagrammatic reasoning task, and use the model's learnt associations between problem and solution states to argue that students are using learnt perceptual chunks as a guide to problem decomposition.

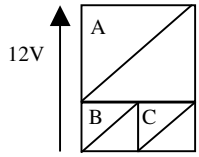
This paper describes an example diagrammatic reasoning task, which involves using AVOW diagrams to compute quantities within circuit diagrams. Based on results from an experimental study, we provide some samples of how students tackle a complex problem within this domain, and observe that the problem is decomposed in a consistent form across the students. The natural question is whether these groupings are based upon the students' prior learning, or are merely an artifact of the target's visual form. We argue that students do learn these groupings, and that their learning process is explainable in terms of the perceptual chunking theory. In support of this, we trained the CHREST+ model, which is based on the perceptual chunking theory, on the same sequence of circuits as was provided to the human participants. We show that CHREST+'s predicted decomposition of the sample problem matches that used by the participants.

Computing Unknowns in Circuits

This paper explores how students learn to use diagrammatic representations for problem solving. As an example domain, we use the task of computing unknown quantities in electric circuits using AVOW (Amps Volts Ohms Watts) diagrams, which represent circuits and the domain laws of electricity using diagrams and constraints on their composition. AVOW diagrams are described in Cheng (1998), and Figure 1 provides an example of how AVOW diagrams are constructed and used for problem solving. Essentially, each resistor in an electric circuit is represented as a separate AVOW box. The dimensions of the box are scaled to represent the quantities within a resistor: current (I) is the width of the box, voltage (V) is the height of the box, resistance (r) is the gradient of the box's diagonal, and power (P) is the box's area. Note that the relations between the box's dimensions encapsulate rules of electric circuits. Thus, the gradient of the box's diagonal is its height divided by its width, which restates Ohm's Law, $r = V/I$. Similarly, the box's area is its height times its width, which restates the Power Law, $P = VI$.



The battery has a voltage of 12V and all three of the resistors have a resistance of 1 ohm. What is the current drawn from the battery?



Each of the AVOW boxes is a square, because each resistor's resistance is 1 ohm. The height of the total diagram is scaled to represent 12V. Hence, the width of the total diagram represents the current, measured to be 8 amps.

Figure 1 : A circuit problem and its AVOW diagram.

Composition of individual boxes is used to represent a circuit of several resistors; the rules for composition preserve the underlying physical laws of electric circuits. In working with this representation, students must first produce an AVOW diagram scaled according to the provided quantities, and the constraints in the diagram ensure that the laws of electricity are followed during its composition. The final AVOW diagram will thus provide information about all other quantities within the circuit, enabling the student to simply measure the appropriate dimension for any unknown quantity. Various studies (Cheng, 1998, 1999a) have shown that AVOW diagrams provide a more effective representation than algebra for learning concepts about electric circuits. An increasingly important element in the design of effective educational material is a better understanding of how humans learn with these representations (Cheng, 1999b). The aim of this paper is to find some indicator in the students' solutions to their underlying learning mechanisms. We achieve this by showing that all participants use a similar decomposition of the circuit problems.

Observing the Problem Decomposition

A study on the use of AVOW diagrams was performed with six participants (2 with A level physics, 4 without). Each participant received basic instruction in the use of AVOW diagrams, and was then asked to construct appropriate AVOW diagrams for a sequence of electric circuits. Solution diagrams were entered on an electronic sketchpad, which allows diagrams to be constructed on screen using a mouse to place elements such as lines, rectangles and parallel lines of various thicknesses, as well as add textual labels. The computer retains a record of each drawing action with detailed timing information; note, the system provides no support for constructing AVOW diagrams *per se*. After an initial 15 minutes' training session on AVOW diagrams and in using the electronic sketchpad, participants were presented with a graded sequence of 30 problems, ending with complex circuits of up to 12 resistors. After each circuit was attempted, the correct AVOW diagram was shown to the participants. We illustrate here how the participants performed on the last of the 'straight-forward' circuits, illustrated in Figure 2(a) with its target AVOW diagram in Figure 2(b). (The remaining four circuits tested for generalisation to more complex circuit types, such as those requiring 3D layouts, and so are not included.)

Figure 4, at the end of this paper, illustrates in detail the progress of three participants on the sample problem. The graphs show the latency between each of the drawing actions required to complete the solution. Noticeable in these examples, and common to all the participants, is the presence of *peaks*, which separate

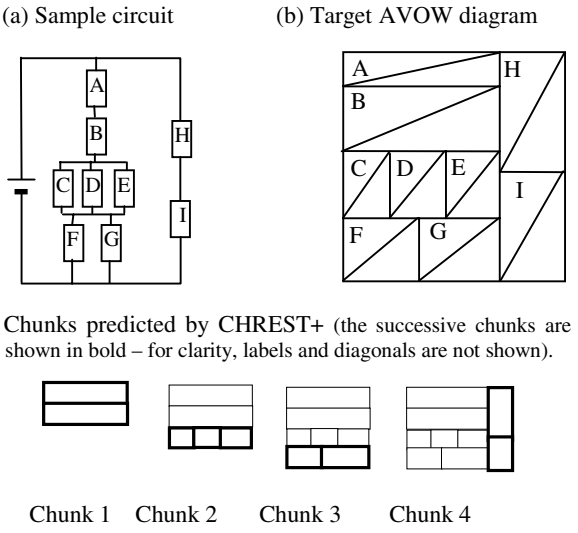


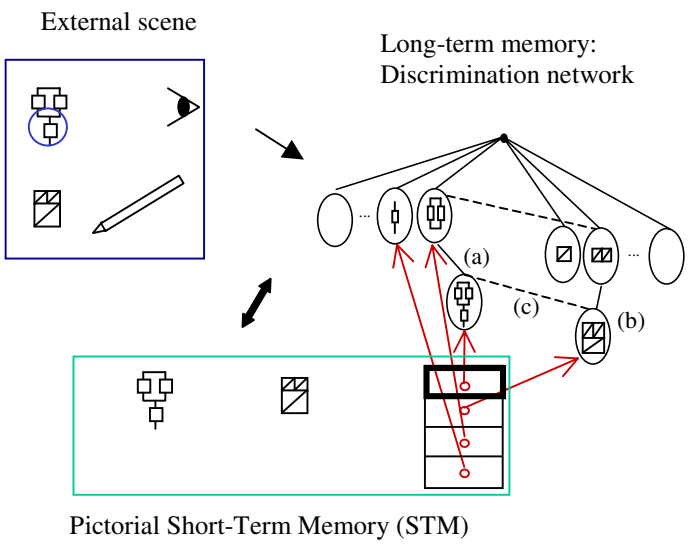
Figure 2 : The sample problem and its solution.

the sequence of actions into a number of stages. Such peaks have also been shown to correlate with meaningful decompositions in other forms of drawing (e.g. Cheng, McFadzean & Copeland, 2001). The figure also illustrates the parts of the AVOW diagram completed during each stage. These stages represent how the participant decomposed the solution.

From this first look at the graphs, we may conclude that the participants are using similar decompositions. These decompositions must be based on features of the target circuit diagram, and the interesting question from the perspective of effective representational design concerns their origin: Are they mere artifacts of the grouping of elements within the circuit, or are they the result of prior training? It is difficult to answer this question directly without some insight into the knowledge which each participant brings to the sample problem. In order to tackle this question, the next section describes a computational model, CHREST+, and shows how it can be used to predict the behavioural characteristics found in the participants' data.

CHREST+ : Learning Perceptual Chunks

The perceptual chunking theory for human memory has had a long history within cognitive science, and forms the theoretical basis of the EPAM/CHREST family of computational models (for a review, see Gobet *et al*., in press). Chase and Simon (1973) first proposed how perceptual chunking could be used in a model of problem solving based on EPAM (Elementary Perceiver and Memorizer) (Feigenbaum & Simon, 1984). The EPAM model assumes an input device (e.g. a simulated eye), a short-term memory (STM) for storing intermediate results, and a long-term memory based around a discrimination network containing



Recognition

The simulated eye is passed over the target stimulus, and the retrieved patterns are sorted through the discrimination network. Matching chunks are retrieved and placed within STM.

Learning

New nodes are added to the network through a process of *discrimination*. Illustrated at (a) and (b), the new pattern has caused an additional test link and node to be created. *Familiarisation* has then filled in the node image with features from the observed pattern. *Equivalence links* (the dotted lines) are formed when CHREST+ has a pointer to both a circuit chunk and an AVOW chunk within its STM at the same time. Here, the model has just learnt link (c).

Figure 3 : The CHREST+ model. The model includes a simulated eye and pen for interacting with the external environment, a fixed capacity short-term memory, and a long-term memory.

chunks of information. However, EPAM itself was only applied to certain perception and memory phenomena, and not more complex problem-solving domains, in part because of the simplified form of its learning mechanisms. This limitation is corrected in the CHREST (Chunk Hierarchy and REtrieval STructures) model, which includes various extensions to EPAM (Gobet, 1996; Gobet & Simon, 2000).

CHREST+ (Lane, Cheng & Gobet, 2000) has been developed to investigate how a memory of perceptual chunks can be used in problem solving with diagrams; the model is illustrated in Figure 3. Like CHREST, CHREST+ learns a discrimination network of perceptual chunks by scanning circuit and AVOW diagrams with its simulated eye. The network consists of a collection of perceptual chunks, which are stored at nodes in a network, interconnected by test links. Patterns are used to retrieve chunks from the network by sorting them, beginning from the root node, through the network against the tests stored at the test links. Once a pattern reaches a node, learning may occur: if the pattern matches the chunk at the node, then more information can be added from the pattern to the chunk (familiarisation); if the pattern mismatches the chunk, then a new test link is added based on the mismatching features, and a new node is created (discrimination). These processes are illustrated in Figure 3: at point (a), a test link for a single resistor is used to distinguish the collection of three resistors from the two in parallel; this link is added during discrimination, and the contents of the node added during familiarisation.

The process of learning about circuit and AVOW diagrams is illustrated in Figure 3. Because the circuit and AVOW diagram representations do not overlap, individual nodes within the network will represent either an individual circuit diagram, or an individual AVOW diagram. In consequence, if CHREST+ is to generate AVOW diagrams when presented with a circuit, it must also associate chunks about circuits with chunks about AVOW diagrams. Accordingly, CHREST+ includes an additional learning mechanism for forming *equivalence links*; these are lateral links (Gobet, 1996) connecting two chunks within the discrimination network. An equivalence link is formed when the model is presented with a circuit diagram and its equivalent AVOW diagram. During the process of recognising the two diagrams, separate chunks will be placed into STM, one for the circuit diagram and one for the AVOW diagram. An equivalence link is then formed between the relevant two nodes in the network. Figure 3 illustrates this process, with an equivalence link formed at point (c). As can be seen, particular problems (circuit diagrams) become associated with information about their solution (equivalent AVOW diagrams). Generating an AVOW diagram for a novel circuit diagram then requires the model to locate chunks (sub-networks) within the circuit diagram for which it has an associated AVOW diagram; the AVOW diagrams for these sub-networks may then be drawn, and a further familiar sub-network located. The process by which CHREST+ incorporates its retrieved AVOW diagram into the evolving solution diagram is provided by specific, handcoded routines – these are akin to the basic training the participants received in AVOW diagrams.

Predicting the Observed Decompositions

We trained CHREST+ using the same sequence of circuits as the participants. By the time CHREST+ reaches the sample problem, it has learnt a discrimination network of 72 chunks, 42 for circuit diagrams and 30 for AVOW diagrams, with 11 equivalence links. When presented with the sample problem, CHREST+ retrieves four separate chunks whilst constructing its solution; these are illustrated in Figure 2(c). Note that the assumption in CHREST+ that information is contained in encapsulated chunks strongly predicts that problems will be decomposed as familiar chunks. Also, because CHREST+'s chunks are associated directly with equivalent AVOW diagrams, we can observe the effect of its circuit decomposition in the breakdown of the AVOW diagram's construction into stages. We now show how the decompositions can be affected by learning, how the participants' data provide reliable decompositions, and how well the participants' data are matched by the model.

Decompositions are Due to Learning

The precise number and content of chunks used by CHREST+ is governed by its experience with the previously encountered problems. By providing different sets of problems, CHREST+ extracts different familiar chunks when decomposing the same sample circuit diagram. For example, presenting the sample problem after initial training on a circuit containing only a single resistor leaves CHREST+ with little choice but to decompose the problem into 9 distinct resistors; a more extensive training sequence allows CHREST+ to identify just 2 sub-circuits within the sample circuit. With the training sequence used, CHREST+ therefore makes a two-fold prediction: that four chunks are identified in the circuit, and that their form is as illustrated in Figure 2(c).

Specifying a Decomposition

Returning to the graphs of the participants' drawing actions shown in Figure 4, we can consider how the decompositions provided by the peaks in the drawing actions compare across participants, and also whether they compare with the model's predicted chunks.

The first participant, CP, in Figure 4(a) has clearly begun from the right-hand side of the figure, then worked out the central triplet of resistors. These stages are preceded by longer pauses between the drawing times (marked by asterisks), and their correspondence to the chunks given by the model is clear: we highlight the stages with vertical divisions, illustrating the current state of the solution at the end of each stage. Similarly with the second participant, SG, who instead begins from the left-hand side; note also that SG requires considerably more time than CP. Note the different order in which the diagram is tackled, although the overall stages are the same. Finally, the third participant, EF, used whole rectangles when constructing the solution; these rectangles were laid out in sequential fashion, beginning from the right-hand side, and then top to bottom. However, from the pauses evident in the times between drawing actions, we can see that this sequence of boxes was divided into the four stages corresponding to individual chunks. The remaining three participants show a similar pattern, but are not illustrated here.

Table 1 : The number of actions each participant made when completing the sample problem, classified as follows: NP – non-peaks; P – peaks; PC – pre-chunk; SB – start+bounding box; L – labels; E – error; Mi – missed chunks; ∑ – totals.

	Participant						
	CP	**DJ**	**EF**	**HA**	**RH**	**SG**	**∑**
NP	19	17	7	16	8	8	75
P	5	5	4	7	3	5	29
PC	3	3	2	3	2	4	17
SB	2		1	1	1	1	6
L		2		3			5
E			1				1
Mi	1	1	1		2		5

Matching Observed Decompositions

We can now directly compare the stage-wise output of the sample circuit's AVOW diagram by CHREST+ with its solution by the participants. We quantify the correspondence between CHREST+'s prediction and the participants' behaviour by counting how many of CHREST+'s chunk boundaries correspond with the participants' peaks. For this analysis, a peak is a time between drawing actions prominently larger than the preceding and succeeding times: the peaks used are highlighted in the figure with asterisks.

For example, the graph for CP shows five peaks. The first and second peaks correspond to CP beginning the problem and creating a bounding box for the entire circuit, as illustrated in the diagram before the first dividing line. Between the third and fourth peaks, CP completes the part of the diagram which corresponds with chunk 4 in CHREST+'s output, and hence we count the third peak as a pre-chunk boundary. Similarly, between the fourth and fifth peaks, chunks 2 and 3 are completed in the diagram. Note that there is no peak corresponding to a retrieval of the 3rd chunk, against the model's prediction. Finally, after the fifth peak, CP completed chunk 1 and then added the labels to all the AVOW boxes; the simplicity of this process is reflected in the low times between these operations.

We therefore explain the five peaks as follows: the first two are for the start and bounding box, and the next three are pre-chunk boundaries. One chunk boundary seems to have been missed. Table 1 summarises the analysis for all six participants.

The analysis shows that nearly all of the peaks correspond with stages in the drawing which we would explain by the use of chunks. Note that, out of the 24 predicted chunks (6 participants and 4 predicted chunks), 17 peaks were clearly identifiable chunk boundaries, and there were only 5 missed chunks. This leaves unaccounted 2 chunks, which were instead created directly from the start, and thus are included with the starting times of the participant: this analysis therefore identifies 19 peaks in the participants' behaviour which precede chunk boundaries out of 24 predicted chunks. These results demonstrate that CHREST+ predicts the decomposition of problems evident in the participants' data, and so support our claim about the role of learnt perceptual chunks in problem decomposition.

Conclusion

This paper has used the perceptual chunking theory, as implemented in CHREST+, to predict specific perceptual chunks learnt from a given sequence of instruction. We have presented results from a study of six participants solving electric-circuit problems using AVOW diagrams. The predictions from the model have been shown to correlate with the stages in problem solving evident in the participants' performance.

To fully understand the participants' problem decomposition and learning pattern, we need to consider more closely what is happening during the peaks in their output timings. Looking at the CHREST+ model, the participants' peaks correspond to the processes of pattern recognition and retrieval. In addition, processes of planning along with some lookahead must be going on. This lookahead and planning probably explains the missed chunks in the preceding analysis. At present, CHREST+ always outputs its solution AVOW diagram as soon as it is found. Through a small modification in its output strategy, CHREST+ could instead retain more than one chunk for solution, and output several together. This would provide CHREST+ with the potential for lookahead, making it a more plausible problem solver, as well as capture the pattern of missed chunks.

The interesting result from this paper is that the chunks used by learners within such domains may be predicted using an established computational model. Further work with this domain should aid in refining the model and its predictions, and also extend it into other domains. In particular, this use of perceptual chunks in decomposing diagrams has already been shown to occur more generally (Cheng, McFadzean & Copeland, 2001). In the longer-term, one of the important applications for this research is likely to be the design of effective computer-based learning environments (Gobet & Wood, 1999).

Acknowledgements

The authors would like to thank Lucy Copeland for conducting the experiments discussed in this paper.

References

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), *Visual Information Processing*. New York: Academic Press.

Cheng, P. C-H. (1996). Scientific discovery with law-encoding diagrams. *Creativity Research Journal*, 9, 145-162.

Cheng, P. C-H. (1998). A framework for scientific reasoning with law encoding diagrams: Analysing protocols to assess its utility. In M. A. Gernsbacher & S. J. Derry (Eds.) *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 232-235). Mahwah, NJ: Erlbaum.

Cheng, P. C-H. (1999a). Electrifying representations for learning: An evaluation of AVOW diagrams for electricity. (Technical Report 62, ESRC Centre for Research in Development, Instruction and Training, University of Nottingham).

Cheng, P. C-H. (1999b). Unlocking conceptual learning in mathematics and science with effective representational systems. *Computers and Education*, 33, 109-130.

Cheng, P. C-H., McFadzean, J. & Copeland, L. (2001). Drawing out the temporal signature of induced perceptual chunks. In *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society* (this volume).

Feigenbaum, E. A., & Simon, H. A. (1984). EPAM-like models of recognition and learning. *Cognitive Science*, 8, 305-336.

Gobet, F. (1996). Discrimination nets, production systems and semantic networks: Elements of a unified framework. In *Proceedings of the Second International Conference of the Learning Sciences* (pp. 398-403). Evanston, III: Northwestern University.

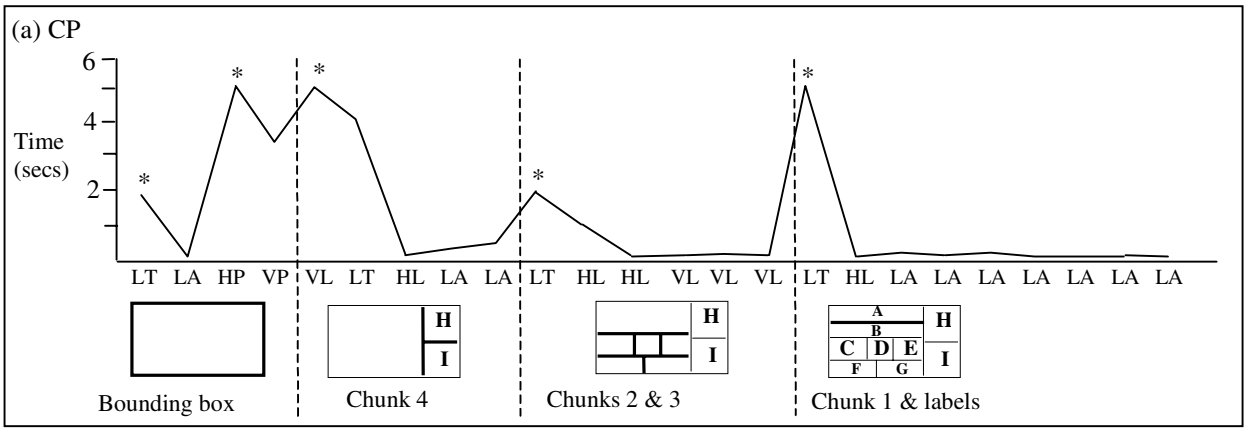
Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C-H., Jones, G., Oliver, I., & Pine, J. M. (in press) Chunking mechanisms in human learning. *Trends in Cognitive Science.*

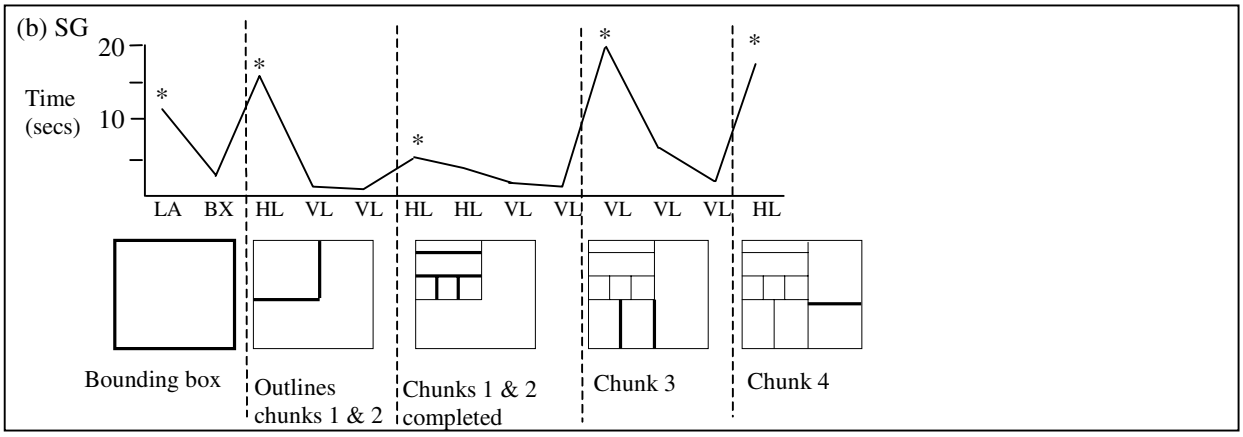
Gobet, F., & Simon, H. A. (2000). Five seconds or sixty? Presentation time in expert memory. *Cognitive Science,* 24, 651-682.

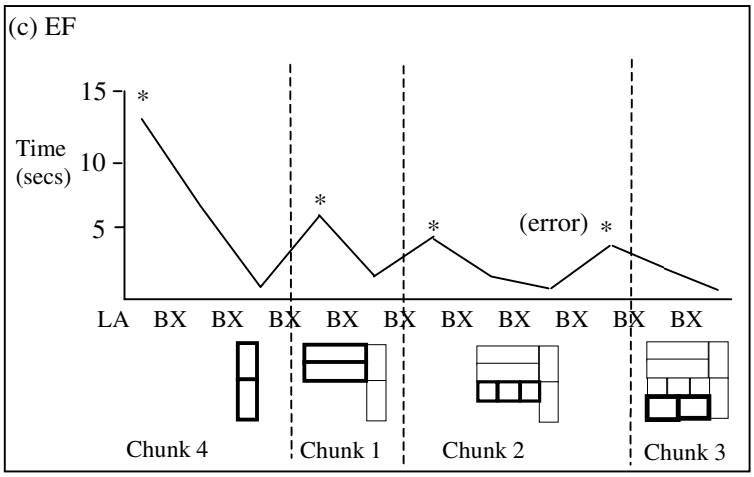
Gobet, F., & Wood, D. J. (1999). Expertise, models of learning and computer-based tutoring. *Computers and Education*, 33, 189-207.

Lane, P. C. R., Cheng, P. C-H., & Gobet, F. (2000). CHREST+: Investigating how humans learn to solve problems using diagrams. *AISB Quarterly*, 103, 24-30.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11, 65-99.







Key to drawing operations:

BX	Box
HL	Horizontal Line
HP	Horizontal Parallel Lines
LA	Add Text Labelling
LT	Change Line Thickness
VL	Vertical Line
VP	Vertical Parallel Lines

Bold lines indicate the features added during each stage.

Figure 4 : The performance of three participants when solving the sample problem. The * indicates peaks used in the analysis. Under the graphs, information is given on the specific drawing action performed, the stage of the diagram just prior to the peaks, and the correspondence with the chunks predicted by CHREST+.

The Mechanics of Associative Change

M.E. Le Pelley (mel22@hermes.cam.ac.uk)
I.P.L. McLaren (iplm2@cus.cam.ac.uk)
Department of Experimental Psychology; Downing Site
Cambridge CB2 3EB, England

Abstract

Rescorla (in press) investigated the change in associative strength undergone by cues A and B as a result of reinforcement or nonreinforcement of an AB compound. Many leading theories of associative learning predict that if A and B are equally salient then the associative change experienced by each should be the same regardless of their associative strength preceding AB trials. Rescorla explored this prediction for a compound composed of an excitatory A and an inhibitory B, using rats and pigeons as subjects. We repeated Rescorla's experiment using human subjects and a causal judgment task, and obtained diametrically opposite results to those of Rescorla's earlier study. The implications of this finding are discussed with reference to a number of influential theories of associative learning.

Introduction

It has long been recognised that stimuli presented in compound can, and will, interact and compete for associative strength. This is powerfully demonstrated in the phenomenon of blocking (Kamin, 1969). This refers to the finding that the gain in excitatory strength accruing to a conditioned stimulus (CS), B, following reinforcement (+) of an AB compound is much reduced if cue A has previously been trained as being a good predictor of that outcome (unconditioned stimulus, US). Learning does not simply progress with each cue independently. Instead the two cues seem to compete for a limited amount of associative strength. Such demonstrations of cue competition provided the motivation for the development of models that can deal with the issue of predictive redundancy in associative learning (e.g. Rescorla & Wagner, 1972; Mackintosh, 1975; Pearce, 1987).

A common feature of these models is the idea that the magnitude of associative change depends in some way on the discrepancy (or *error*) between the current associative strength of the presented cues and the strength which the outcome (unconditioned stimulus, US) following these cues can support. Consider, for instance, the Rescorla-Wagner (1972) model (R-W), perhaps the most influential of all of these "error-correcting" theories:

$$\Delta V_A = \alpha_A \beta_{US} (\lambda_{US} - \Sigma V) \qquad (1)$$

where ΔV_A is the change in associative strength of cue A, α_A and β_{US} are rate parameters relating to the salience of cue A and the US respectively, λ_{US} is the asymptote of conditioning supportable by that US, and ΣV is the summed associative strength of all cues present on a trial. Hence R-W states that the error governing associative change for any cue on a trial is based on the combined associative strength of *all* cues present on that trial. This is essential to R-W's explanation of blocking. On A+ trials, V_A will increase, with A coming to predict the US. On AB+ trials, the error term for cue B (and also for A) will be $(\lambda - \{V_A + V_B\})$. But of course as a result of A+ training V_A will already be high, and so the error term will be correspondingly small, such that any increase in V_B will be only very small. Thus the R-W explanation of blocking crucially hinges on the idea that, when determining the associative change undergone by B, the associative strength of other cues present on the trial (A) is also considered.

This use of a common error term governing associative change for all stimuli on a trial has important consequences. Rescorla (in press) noted that, in the absence of additional assumptions, it predicts that equally salient stimuli presented together on a trial will undergo equal associative changes. This prediction holds true regardless of the associative history of the cues in question.

In a recent series of experiments, Rescorla (in press) investigated this prediction in rats and pigeons (using magazine approach conditioning and autoshaping procedures respectively). He looked at the particular instance of an AB compound composed of an excitatory A and an inhibitory B. Specifically, he was interested in the associative change undergone by A and B as a result of either reinforcement or nonreinforcement of the AB compound. If we assume that A and B are of equal salience (ensured by counterbalancing) then, as a result of using a common error term, R-W is constrained to predict that both A and B will show equal associative change following either AB+ or AB- trials. Consider, for example, the AB+ condition. If A and B are equally salient, then $\alpha_A = \alpha_B$. Since both are presented with the same US, β will also be equal when calculating ΔV_A and ΔV_B according to equation (1). And finally, the error term for both A and B will be $(\lambda - \{V_A + V_B\})$. Hence, given that all the terms in the calculation for ΔV_A and ΔV_B are identical, Rescorla-Wagner must predict that on AB+ trials, $\Delta V_A = \Delta V_B$. This prediction of equal associative *change* holds true despite the fact that A (an excitor) and B (an inhibitor) begin these trials with very different associative strengths ($V_A > 0$, $V_B < 0$).

The problem with investigations such as this is one of how to assess the magnitude of associative change for two stimuli that differ in their "baseline" associative strength. It would be unwise to make any strong assumptions with regard to mappings between associative strength and measurable performance, and yet without such mappings we cannot be sure that two equal-sized changes in performance at different points on the performance scale represent equal-sized changes in associative strength.

Condition	Stage 1		Stage 2		Test
CR	A+	C+	AB+		**AD**
	E+		CD?		**BC**
	BE-	DE-			
CNR	F+	H+	FG-		**FI**
	J+		HI?		**GH**
	GJ-	IJ-			
Fillers	KL+	MN+	K-	M-	
	OP+	Q-	Q-	V+	
	R-	S-	W+	X+	
	T-	U-			

Table 1. Experimental design.
+: outcome; -: no outcome; ?: exposure trial.

Rescorla suggested an elegant way of avoiding this problem, by comparing performance to A and B when they were embedded in compounds designed to ensure comparable overall levels of performance. We adopt this technique in our experimental design (Table 1). Consider the Compound Reinforcement (CR) condition. A and C are initially trained as equivalent excitors, while B and D are trained as equivalent inhibitors. So following Stage 1, compounds AD and BC should have equal strengths, as each contains one of two equal excitors and one of two equal inhibitors. We then reinforce the AB compound. If this results in equal changes to the associative strengths of A and B (as predicted by R-W), then the AD and BC compounds should remain equal after Stage 2 (as each starts at the same level and receives the same change). If instead the strength of A increases more than that of B, then responding to AD will be greater than to BC. Conversely, if the strength of the inhibitory B increases more than that of the excitatory A, then BC will give rise to more conditioned responding than AD. A similar argument can be applied to the Compound Nonreinforcement (CNR) condition: if FG- trials cause a greater decrement in V_F than V_G then we expect the FI compound to be rated lower than GH: if V_G decreases more than V_F we expect the opposite.

Using this kind of logic, the results of Rescorla's experiments indicated that reinforcement of AB led to a greater increase in the associative strength of the inhibitory B than the excitatory A, while nonreinforcement of AB led to greater associative loss in the excitatory A than the inhibitory B. Rescorla (2001) carried out a similar investigation, this time with an AB compound consisting of an excitatory A and a neutral B. Again, AB+ trials led to a greater increase in V_B than V_A, whereas nonreinforcement gave a greater decrement in V_A than V_B. This indicated that the previous result was not simply due to some special property of conditioned inhibitors. Instead it seems that initial associative strength is an important factor in determining the distribution of associative change among the elements of a compound. This, of course, runs contrary to the predictions of R-W.

This prediction of equal change does not apply only to elemental theories such as R-W. Consider Pearce's (1987) configural model. This model proposes that a compound stimulus is best viewed as a unitary event that is separate from its elements, but able to generalise to them. Whereas according to an elemental model, an AB compound is decomposed into separate A and B elements, in a configural model it is represented as a single "AB" configuration. Generalised responding to other stimuli (e.g. A alone or B alone) occurs to the extent that these stimuli are similar to previously experienced configurations.

So on AB+ trials in Stage 2 of the CR condition, a configural model such as Pearce's learns an association between an AB configural unit and the US. Assuming that A and B are equally salient, this excitatory learning will generalise equally to each of them (they each have the same degree of similarity to AB), and so the model is constrained to predict equal associative change for A and B as a result of AB+ trials. Again no reference is made to the associative history of the cues. A similar story applies to the CNR condition: any change in association from an FG configural unit to the US will generalise equally to F and G.

Dickinson, Shanks & Evenden (1984) noted many similarities between Pavlovian conditioning in animals and the acquisition of causal judgments in human subjects. However, Le Pelley & McLaren (in press) demonstrated that not all phenomena in the animal learning field have analogues in studies of human causal learning. Given the importance of the previously described findings in elucidating the mechanisms underlying associative change, we aimed to repeat Rescorla's (in press) experiment using a causal judgment task with human subjects. This is of particular interest to us as, if we were to replicate Rescorla's findings in human subjects, it would invalidate the APECS model of associative learning that we have developed in recent years (Le Pelley & McLaren, in press; Le Pelley, Cutler & McLaren, 2000; see also McLaren 1993, 1994).

APECS is a model of learning and memory, based on the popular backpropagation algorithm (Rumelhart, Hinton & Williams, 1986), but with a couple of important differences. Firstly, APECS employs configural representation. Thus each different mapping of input to output is represented by its own hidden unit, which could equally well be termed "configural units". Secondly, APECS uses adaptive generalisation coefficients to determine the amount of generalisation between similar input patterns. As a result, once the weights appropriate to a mapping have developed, the learning in those weights can be protected against interference. This is achieved by reducing the learning rate parameter for the configural unit carrying that mapping. The effect is to "freeze" the weights to and from a certain configural unit at the value they hold immediately following experience of that configuration. Crucially, this freezing of weights to and from a certain configural unit occurs only if that configural unit has a negative error value, i.e. *if it is part of a mapping that predicts an incorrect outcome for the current input*. APECS has different learning rate parameters for input–hidden and bias–hidden connections. The former are frozen to prevent interference; the latter remain high. Hence extinction (suppression of inappropriate responses) is achieved by an increase in the negative bias on the hidden unit carrying the inappropriate mapping, rather than by reduction of weights (which would cause the original mapping to be lost from the network). Given appropriate input cues, the negative bias on the hidden unit can be overcome and the original mapping retrieved. So bias acts to change the retrievability of previously learnt mappings, such that

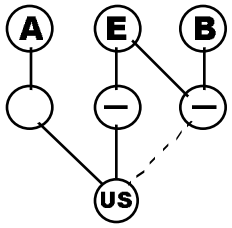


Figure 1. Associations developed following Stage 1, according to APECS. Excitatory connections are shown by solid lines, inhibitory associations by dotted lines. Negative bias on hidden units is indicated by a minus sign.

APECS addresses both learning (in formation of weights) and memory (in changes of retrievability).

Consider the processes at play in the CR condition, according to APECS. During Stage 1, the network will learn an excitatory connection from a representation of A to a configural unit, and from the configural unit to the output. Hence this configural unit comes to represent the A+ mapping. A different configural unit will be recruited to carry the E+ mapping. On BE- trials, presentation of E will cause positive activation to flow to the output via this E+ excitatory pathway. This positive activation is inappropriate on a trial on which the US is not presented: as a result the output unit will take on a negative error. This negative error will be propagated back along the excitatory connection to the E+ configural unit. This configural unit will therefore take on a negative error, as it is part of a mapping predicting an inappropriate outcome for the current input. This negative error means that, on BE- trials, the weights to and from the E+ configural unit will be frozen, as stated above. Instead the E+ configural unit will take on a negative bias to reduce expression of this excitatory mapping on these nonreinforced trials. In addition, excitatory associations will develop from B and E to a new configural unit, representing the BE configuration. This unit will develop an inhibitory association to the output in order to further counter any positive activation flowing to the output via the E+ hidden unit. By a similar argument, this BE- configural unit will develop negative bias on E+ trials. Hence following Stage 1, the situation for cues A, B and E is as shown in Figure 1 (this also applies to cues C, D, F, G, H and I and J, all of which have an equivalent partner in A, B or E following Stage 1).

What now happens on AB+ trials? The US will receive some positive activation via the A+ mapping learnt in Stage 1. However, it will also receive some negative activation via the BE- mapping (which can never be *totally* suppressed by development of negative bias). As a result the US will not be perfectly predicted on these trials, and yet is presented. Therefore the output unit will have a positive error. How can this error be reduced? Well, the positive error on the US unit will be propagated back to the BE- configural unit. But it is propagated along a negative connection, and so the BE- unit will take on a negative error (again, it is part of a mapping predicting an incorrect output on this trial). Thus weights to and from this unit are frozen. Extra negative bias can still be applied to the unit to reduce the negative activation flowing to the output, though, and this will help to reduce the output error. The positive output error will also be propagated back to the A+ configural unit along the positive connection. Thus the A+ configural unit will have a positive error. Both its weights and its bias are therefore free to change: the connections from A input to A+ configural, and from A+ configural to output, will increase. This too will reduce the output error.

In our previous expositions of APECS, we have always made the assumption that changes in weights occur faster than changes in bias. Thus we assume that changes due to learning take place faster than changes in memory, i.e. that learning represents rapid acquisition, and memory represents a more gradual decline in retrievability: this seems reasonable. We saw above that on AB+ trials, the weights of the A+ mapping are free to increase, whereas only the bias of the BE- mapping may change to reduce the effective strength of the inhibitory mapping. Therefore APECS is constrained to predict that, on these AB+ trials, the associative strength of the excitatory A will increase more than that of the inhibitory B. This is of course opposite to Rescorla's result.

A similar argument holds for the CNR condition. On FG- trials, the F+ hidden unit will have negative error (so only its bias may change), while the JG- unit has positive error (so that its weights and bias may both change). In this case APECS must predict that, on FG- trials, the associative strength of the inhibitory G will decrease more than that of the excitatory F, again opposite to Rescorla's result.

In summary, these results follow from the idea of adaptive generalisation. AB+ training generalises more to A+ than it does to BE- because AB and A predict the same outcome. Similarly FG- learning generalises more to GJ- than to F+ as FG and GJ predict the same outcome (no US).

Rescorla noted a problem with his paradigm that could cast doubt on the results obtained. AB compound presentation may result not only in development of A–US and B–US associations, but also in development of within-compound A–B associations. Consideration of these A–B associations complicates any inference of unequal associative change drawn from the results. Suppose A and B undergo equal changes as a result of AB+ trials. The formation of an A–B association might be expected to enhance responding to the inhibitory B (as it has been paired with an excitor), and reduce responding to the excitatory A (as it has been paired with an inhibitor). Thus even if the change in the A–US and B–US associations were equal, one would expect that AB+ trials would augment responding to B more than to A. Similarly, nonreinforcement of the AB compound might result in equal A–US and B–US decrements, but responding to A may fall further as it forms an association to the inhibitory B.

Rescorla controlled for the effect of within-compound associations in his Experiments 5 and 6. His findings were unchanged: Stage 2 AB+ trials gave a greater change in B than A, and *vice versa* for Stage 2 AB- trials. We were also careful to control for the effect of within-compound associations. In Stage 2, in addition to AB+ trials subjects also experienced CD "exposure trials". On these trials subjects saw cues C and D paired, but were not told whether or not the outcome occurred. Hence on these trials within-compound C–D–associations would form while C–US and D–US associations remain unchanged. The effect of A–B association formation would thus be matched by development of C–D associations. Therefore any difference between

AD and BC following Stage 2 could only be due to unequal changes in A–US and B–US associations on AB+ trials. The same holds true for the CNR condition.

Our investigation used an allergy prediction paradigm with human subjects. This paradigm has been used successfully in several studies of human causal learning (e.g. Dickinson & Burke, 1996; Le Pelley, Cutler & McLaren, 2000). Participants play the role of a food allergist judging the likelihood that various foods will cause an allergic reaction in a hypothetical patient. The foods, then, constitute the cues; the allergic reaction is the US. Following training, subjects rated how strongly certain individual foods, and compounds of two foods, predicted the occurrence of an allergic reaction. These ratings were taken as our measure of the strength of conditioning. This was a within-subjects experiment: subjects experienced all the different contingencies concurrently.

The Filler trials were included to ensure equal numbers of positive and negative trial types in each stage. In addition, they increase the number of different trial types seen by subjects, again following Dickinson & Burke and Le Pelley et al. This creates a large memory load, hopefully preventing subjects from basing their ratings on inferences made from explicit episodic memories of the various trial types. Instead subjects should have to rely on associative processes to provide an "automatic" measure of the causal efficacy for each cue. Using a large number of trial types makes us more confident that it is indeed associative, rather than cognitive, processes being tapped in our study.

Method

Participants Twenty members of Cambridge University (10 female, 10 male; age 19-49) took part in the experiment.

Procedure At the start of the experiment each subject was given a sheet of instructions presenting the "allergy prediction" cover story for the experiment. They were told that in the first block they would arrange for Mr. X to eat different meals on each day, and would monitor whether he had an allergic reaction or not as a result. In relation to the exposure trials (that do not bear on the issue at hand in this paper), subjects were told that occasionally the results of eating the foods had been lost. On these trials they would know the foods eaten in the meal, but not the result of eating those foods. They were also told that later on they would be asked to rate some of the foods according to how strongly they predicted allergic reactions.

On each conditioning trial, the words "Meal [meal number] contains the following foods:" followed by the two foods appeared on the screen. Subjects were then asked to predict whether or not eating the foods would cause Mr. X to have an allergic reaction, using the "x" and "." keys (counterbalanced). The screen then cleared, and immediate feedback was provided. On positive trials the message "ALLERGIC REACTION!" appeared on the screen; on negative trials the message "No Reaction" appeared. If an incorrect prediction was made, the computer beeped. On the exposure trials of Stage 2, the same message appeared, but now subjects were cued to enter the initial two letters of each of the foods. This was to ensure that they paid attention to the pairings of foods when no allergy prediction was required. The 24 foods used were randomly assigned to the letters A to X in the experimental design for each subject.

As shown in Table 1, there were 16 trial types in Stage 1, and 8 in Stage 2. Stages 1 and 2 were split into 8 sub-blocks, with each trial type appearing once in each sub-block (hence subjects saw each trial type 8 times). The order of trials within each sub-block was randomised, as was the order of presentation on the screen (first/second) within each compound pair.

After Stage 1, subjects were asked to rate their opinions of the effect of eating certain foods on a scale from -10 to +10. They were to use +10 if the food was very likely to cause an allergic reaction in Mr. X, -10 if the food was very likely to prevent the occurrence of allergic reactions which other foods were capable of causing, and 0 if eating the food had no effect on Mr. X (i.e. it neither caused nor prevented allergic reactions). For clarification, participants also had access to a card on which the instructions on how to use the rating scale were printed. Once a meal had been rated it disappeared from the screen and the next appeared: participants could not revise their opinions upon seeing later meals. Subjects were given a second rating test after Stage 2, when they were asked to rate meals containing either one food or two. Exactly the same test procedure was used.

Results and Discussion

The results of Test 1 indicated that we had been successful in generating conditioned excitation (to A and F, mean rating 8.63) and inhibition (to B and G, mean rating –8.1), as compared to Q (mean rating 0.6), which is never paired with the US and hence should remain neutral. Planned comparisons revealed that the average of A and F (which are equivalent) was significantly higher than Q, which was in turn significantly higher than B and G (which are also equivalent) [$F(1,19)=65.9$ and 93.2 respectively, ps<0.001].

Figure 2 shows the mean rating of the casual efficacy of each of the meals of interest as judged in the test following Stage 2. We see that the compound AD is rated higher than BC. This is confirmed statistically [$F(1,19)=5.87$, $p<0.05$]. This implies that reinforcing the AB compound led to a greater increase in the associative strength of the excitatory A than the inhibitory B. This is, of course, diametrically opposite to Rescorla's earlier findings with rats and pigeons.

In addition, we see that FI is rated significantly higher than GH [$F(1,19)=4.43$, $p<0.05$]. This implies that nonreinforcement of the FG compound led to a greater decrement in the associative strength of the inhibitory G than the excitatory F. Again, this is opposite to Rescorla's findings.

Like Rescorla's, our results suggest that the distribution of associative changes among the elements of a compound depends on the associative history of those elements. This asymmetry in associative change contradicts any model employing a common error term governing associative change

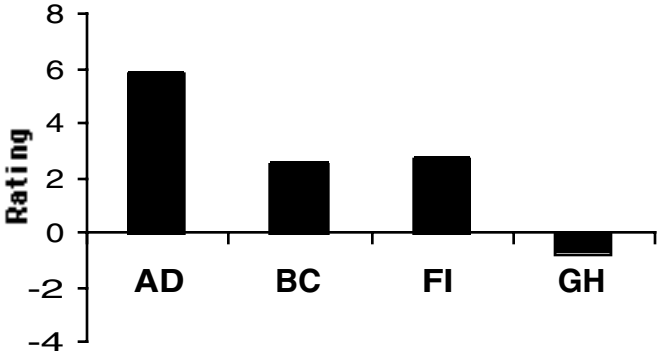


Figure 2. Mean ratings given to the cues of interest.

for all stimuli present, as these models are constrained to predict equal changes for these stimuli.

However, unlike Rescorla's results, we found that it is the cue whose associative strength is *less* discrepant from that supported by the outcome of the trial that undergoes the greater associative change. Our empirical findings agree with the predictions of the APECS model of learning and memory outlined earlier. This was confirmed by simulation. We performed 20 simulations with APECS, each representing a different subject. Each trial involved 1000 learning cycles. A hidden unit is defined as being "active" when it receives positive activation from the input layer. Thus if cue A is presented to the network, any hidden unit representing a configuration that includes cue A will be active. Activity extends into the period immediately following each trial, when no inputs are presented (again for 1000 learning cycles). The learning rate parameters for input–hidden and hidden–output units are both 0.8 when a hidden unit is active and has a positive error, and 0 when it is not. The parameter for bias–hidden changes is 0.3 when a hidden unit is active, 0 when it is not. The exact values of these parameters are unimportant: the pattern of results is robust under quite large variations in the values used. The results of the simulation are shown in Figure 3. As expected, we see that AD is rated higher than BC, and FI is rated higher than GH. Both differences are significant [$F(1,19)=20.2$ and 3.2 respectively, $ps<0.05$]. This is, of course, the pattern seen in our empirical data, and the opposite of Rescorla's results.

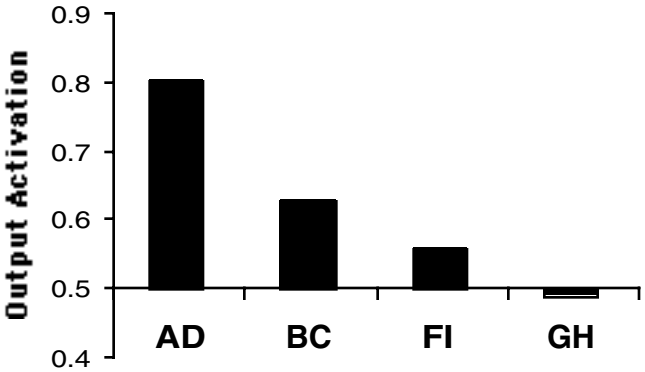


Figure 3. Simulation of the data using APECS.

The theories explicitly considered thus far describe cue competition effects in terms of modifications in the effectiveness of the US. If the US is surprising (i.e. error is high) then it is able to support more learning than if it is already predicted (and therefore less surprising). The contribution of the CS to learning is assumed to be fixed, and is determined by its salience (α, or the parameter for learning weights in APECS). In addition to such US-centred views, there exist a number of influential theories of associative learning that instead ascribe cue competition to variations in the processing of the CS. For example, blocking of B on AB+ trials following A+ pretraining might be interpreted as reduced processing of B as a result of earlier learning about A's predictive power. Typically, these CS-processing models specify a role for attentional processes in determining the distribution of associative change undergone by the cues of a compound. The attention paid to a cue depends on the associative history of that cue, so perhaps these theories would be better-suited to explaining our results.

Mackintosh (1975) proposed just such a model of selective attention. This theory states that good predictors of an outcome will retain a higher salience (i.e. will receive greater attention) than poorer predictors. The calculation determining the attention to be paid to stimulus A relies on a comparison of the predictive power of A *for the outcome occurring on that trial* with the predictive power of all other presented cues for that same outcome. This calculation is carried out after every trial. If cue A is a better predictor of the outcome of that trial than any other cue present, its salience increases for the next trial, and *vice versa*. Hence this model proposes that CSs followed by their expected outcomes (be this reinforcement or nonreinforcement) garner greater salience. CSs followed by surprising events (again, be this reinforcement or omission of reinforcement) lose salience.

According to the Mackintosh theory, learning about each element of an AB compound is governed by the discrepancy between λ and the *individual strength of that element* (rather than the discrepancy between λ and the summed strengths of A and B), modulated by the attention it receives. Thus:

$$\Delta V_A = \alpha_A \beta_{US} (\lambda_{US} - V_A) \qquad (2)$$

where α_A represents the attention paid to cue A.

Consider the CR Condition of our experiment. During Stage 1, A is consistently followed by the US, and B is consistently followed by no US. Thus both will begin Stage 2 with fairly high salience, as they are both good predictors of their respective "outcomes" (which for B is actually nonreinforcement). On Stage 2 AB+ trials, however, A is a better predictor of the outcome (reinforcement) than B, which predicts nonreinforcement. Hence attention to A will remain high, while that for B will be reduced rapidly: increments in V_A will remain relatively high over Stage 2 trials, while increments in V_B will become progressively smaller. As a result, Mackintosh (1975) is able to predict that over all Stage 2 trials, the increment in V_A will be greater than that for V_B. This is, of course, exactly the pattern seen in our empirical data. A similar story holds for the CNR contingency – on Stage 2 FG- trials, the inhibitory G is a better predictor of nonreinforcement than the excitatory F. Hence attention to G will remain high over Stage 2, while attention to F will fall. So overall we might expect a greater decrement in responding to G than to F.

In general, then, and in agreement with our data, Mackintosh (1975) is able to predict that the stimulus whose associative strength is less discrepant from the outcome of the trial (i.e. the better predictor of the outcome) will show the greater change on compound training.

Intriguingly Mackintosh's (1975) can also explain Rescorla's empirical data, which are diametrically opposed to our own, by appealing to the notion of overtraining (Mackintosh, personal communication). If we train subjects on Stage 1 until the associative strengths of excitors and inhibitors closely approach their asymptotic values, then the predictions made by the theory change dramatically.

As a result of this overtraining, A and B will also have very high salience (near asymptote) at the start of Stage 2.

On the initial AB+ trial, then, both will be well processed (as the calculation to update the salience of a cue is performed *after* each trial). Given that the error term governing associative change for a cue involves only the current associative strength of that cue, rather than the summed strength, the stimulus whose associative strength is more discrepant from that supportable by the outcome of the trial will undergo greater change. In other words, on the initial AB+ trial, it will actually be the *poorer* predictor of the trial's outcome (B) that undergoes the greatest associative change, as this cue will have the greater error term. Notably, if A's associative strength is near asymptote (λ), its error term according to equation (2) will be near zero. Given that it is error that drives changes in associative strength, this means that any change in V_A will be only very slight. Of course the modulation of attention discussed earlier will still occur. Thus following this initial trial, attention to A (a good predictor of the outcome) remains high, while that for B (a poor predictor of the outcome) will be reduced. So subsequent changes in V_B will become increasingly smaller. However, given that V_A was already near asymptote at the start of Stage 2, it will undergo little further increase over Stage 2 trials. In other words, the effect of Stage 1 training outweighs any influence of attentional modulation in Stage 2. In fact, the effect of attentional modulation may be reduced even further in the case of an overtrained contingency if we follow Sutherland & Mackintosh's (1971, p. 491) suggestion that the high attentional strengths developed as a result of overtraining are "sticky": a high α value is reduced more slowly than an intermediate value. As such the high value of α_B will persist over several AB+ trials. This, combined with B's high error value on these trials, will result in large increments in V_B over several trials before more significant reductions in α_B start to take their toll on the size of the increments. Thus by appealing to overtraining in Stage 1, Mackintosh (1975) is able to predicts Rescorla's finding that V_B increases more than V_A as a result of AB+ trials.

On this analysis, then, the difference between Rescorla's experiment and our own is that in the former, Stage 1 conditioning led to near-asymptotic associative strengths such that any effect of selective attention in Stage 2 was outweighed. The notion of sticky α values developed as a result of overtraining will further reduce the influence of attentional processes. In our experiment, however, we must assume that Stage 1 training did not approach asymptotic levels, such that both cues in the Stage 2 compound were free to undergo associative change as dictated by their salience.

Conclusion

In common with Rescorla's earlier experiments, our results indicate that a cue's associative history is important when determining the magnitude of its associative change. Unlike Rescorla's experiments, however, our data indicate that it is the better predictor of an outcome that undergoes the greater associative change on compound conditioning. This finding may reflect different rules governing the distribution of associative change among elements of a compound in humans and animals, or may simply be a result of different levels of initial training in the human and animal studies. Given only the results of the current experiment we cannot choose between a "US processing" model of learning and memory employing adaptive generalisation with configural representation (APECS), and a model of selective attention in which CSs compete for attention (Mackintosh, 1975). However, taken in conjunction with several other findings from this laboratory (Le Pelley, Cutler & McLaren, 2000, Le Pelley & McLaren, in press; Le Pelley & McLaren, this issue), we believe that the results of all the human data may be better explained by APECS.

References

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgments. *Quarterly Journal of Experimental Psychology, 49B,* 60-80.

Dickinson, A., Shanks, D. R., & Evenden, J. (1984). Judgment of act-outcome contingency: The role of selective attribution. *Quarterly Journal of Experimental Psychology, 36A,* 29-50.

Kamin, L.J. (1969). Selective association and conditioning. In N.J. Mackintosh & W.K. Honig (Eds.), *Fundamental issues in associative learning* (pp. 42-64). Halifax: Dalhousie University Press.

Le Pelley, M.E., Cutler, D.L., & McLaren, I.P.L (2000). Retrospective effects in human causality judgment. *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 782-787). Hillsdale, NJ: Lawrence Erlbaum Associates.

Le Pelley, M.E., & McLaren, I.P.L. (this issue). Representation and generalization in associative systems.

Le Pelley, M.E., & McLaren, I.P.L. (in press). Retrospective revaluation in humans: Learning or memory? *Quarterly Journal of Experimental Psychology B*.

Mackintosh, N.J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82,* 276-298.

McLaren, I. P. L. (1993). APECS: A solution to the sequential learning problem. *Proceedings of the XVth Annual Convention of the Cognitive Science Society* (pp. 717-722). Hillsdale, NJ: Lawrence Erlbaum Associates.

McLaren, I. P. L. (1994). Representation development in associative systems. In J.A. Hogan & J.J. Bolhuis (Eds.), *Causal mechanisms of behavioural development* (pp. 377-402). Cambridge: Cambridge University Press.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review, 94,* (61-73).

Rescorla, R.A. (2001). Unequal associative changes when excitors and neutral stimuli are conditioned in compound. *Quarterly Journal of Experimental Psychology, 54B*, 53-68.

Rescorla, R.A. (in press). Associative changes in excitors and inhibitors differ when they are conditioned in compound. *Journal of Experimental Psychology: Animal Behaviour Processes*.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black & W.F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart, J.L. McLelland & the PDP Research Group (Eds.), *Parallel Distributed Processing* (Vol. 1, pp. 318-362). Cambridge, MA: MIT Press.

Sutherland, N.S., & Mackintosh, N.J. (1971). *Mechanisms of animal discrimination learning*. New York: Academic Press.

Representation and Generalisation in Associative Systems

M.E. Le Pelley (mel22@hermes.cam.ac.uk)
I.P.L. McLaren (iplm2@cus.cam.ac.uk)
Department of Experimental Psychology; Downing Site
Cambridge CB2 3EB, England

Abstract

This paper examines the nature of stimulus representation in associative learning systems. Specifically, it addresses the issue of whether representation is elemental or configural in nature. We use a human causal learning paradigm, employing contingencies more commonly associated with studies of retrospective revaluation. Whereas most models of retrospective revaluation view it as an entirely elemental process, our results show that it has a configural component. However, the results also prove troublesome for simple configural theories employing fixed generalisation coefficients. It is possible to explain the data using an elemental theory employing configural representation. Our favoured explanation, however, involves a configural theory employing adaptive generalisation. We present such a theory, APECS, and show through simulation that it is well-equipped to deal with our findings.

Introduction

Recent years have seen a great deal of debate concerning the nature of stimulus representation in associative learning systems – more specifically, over how stimulus *compounds* should be represented, and how generalisation between similar compounds should be dealt with. Consider, for example, rewarding the compound AB. Elemental theories (e.g. Rescorla & Wagner, 1972; Mackintosh, 1975; Wagner, 1981) propose that such compounds are represented as being comprised of separable A and B elements that gain individual associative strengths. The conditioned responding shown to a particular stimulus compound is then found by simply adding together the individual associative strengths of each of the elements of that compound.

We sought to test this fundamental assumption of elemental theories, using a causal judgment procedure with human subjects. Our experimental design is shown in Table 1. We used an allergy prediction paradigm – participants play a food allergist trying to judge the likelihood that various foods will cause an allergic reaction in a fictional patient. The foods, then, constitute the cues; the allergic reaction is the outcome. Following training, subjects rate how strongly certain individual foods, and compounds of two foods, predict the occurrence of an allergic reaction. These ratings are taken as our measure of associative strength.

We train two stimulus compounds, AB and CD, to be predictors of the outcome, and then in Stage 2 one of the elements of each compound is extinguished. Cues A and C have an identical training history, as do B and D. The question is, what effect does this treatment have on the originally experienced compounds (AB and CD) as opposed to "elementally equivalent" compounds made up of identically trained cues that have never been seen in compound before (AD and BC)?

An elemental theory predicts no difference between the two types of compound. If the associative strength of a compound is given by adding together the strengths of all the separate elements contained in that compound, then whether or not those elements have been seen in compound before should have no effect. The Rescorla-Wagner (1972) model (R-W), for example, states that:

$$\Delta V_A = \alpha_A \beta_{US} \left(\lambda - \sum V \right) \qquad (1)$$

where ΔV_A is the change in associative strength of cue A, α_A represents the salience of cue A, β_{US} represents the salience of the US, λ relates to whether the US is actually present on a trial (taking a positive value if the US is present, 0 if it is not), and ΣV is the summed associative strength of all cues present on a trial. According to R-W, following Stage 1 all of cues A to F will have associative strengths of 0.5λ (ignoring the effect of α and β, which will be equivalent for all the different cues as a result of counterbalancing). In Stage 2, extinction trials will reduce V_A and V_C to 0. According to an elemental rule, the associative strength of a compound is found by summing the associative strengths of all of the elements of that compound. The associative strength of AB will be given by the total of the associative strength of A (0) plus the strength of B (0.5λ), i.e. 0.5λ. Of course, the compounds BC and AD are also both made up of one element with a strength of 0, and the other with a strength of 0.5λ, and so all of the compounds AB, CD, BC and AD should give rise to the same level of conditioned responding, as they are all elementally equivalent.

Subjects also received EF+ trials in Stage 1, with no further training of either cue in Stage 2. Given that neither E nor F is experienced in Stage 2, the associative strength of EF should remain at λ (as $V_E=V_F=0.5\lambda$). Hence R-W predicts that EF should receive a higher rating than the other

Stage 1			**Stage 2**	
AB+ **CD+** **EF+**			**A-** **C-**	
G+	H+	I+	GL+	Q-
J+	K+	L+	IO-	V+
KM-	KN-	LO-	HJ?	W+
LP-	Q-	R-	JP?	X+
S-	T-	U-		

Table 1. Experimental design. Important trials in bold. +: outcome; -: no outcome; ?: exposure trial.

compounds, which should all receive similar ratings.

However, it is important to note that the AB+, A- design we are using is more commonly associated with studies of retrospective revaluation (see Le Pelley & McLaren, in press, for a review). This term is used to describe changes in the associative status of previously trained cues in the absence of those cues. For example, it is typically found that A- trials following AB+ training lead to an increase in the causal efficacy of B, even though B itself is absent on these trials. This is the phenomenon of unovershadowing, and would be revealed in our experiment by higher ratings given to B and D than to E and F, which receive no such revaluation in Stage 2.

Findings of retrospective revaluation are problematic for many theories of associative learning. R-W, for example, states that α, the salience of a cue, is positive for a cue that is actually presented on a trial, and zero for all absent cues. Hence the theory incorrectly predicts that there will be no learning about absent cues. So V_B remains unchanged at 0.5λ during Stage 2, with R-W thus constrained to predict that B, D, E and F will all receive similar ratings on test.

It is possible, however, to adapt R-W to allow it to predict unovershadowing. Van Hamme & Wasserman (1994) proposed that absent cues, rather than having $\alpha=0$, should take on a negative value of α, thus engaging the learning process with a negative sign. So on Stage 2 A- trials, while A's association to the outcome becomes weaker, the association from the absent cue B to the outcome will become correspondingly stronger. Markman (1989) proposed that only absent *but expected* cues should take on negative α. Dickinson & Burke (1996) suggested that this expectancy arises as a result of within-compound associations formed during Stage 1 compound training. During AB+ trials, subjects learn not only that A and B predict the US, but also that A predicts the presence of B, and *vice versa*. Presentation of A on A- trials now creates an expectancy of the absent cue B, and it is this expectancy that imbues it with negative α.

Modified R-W now predicts that B will be rated higher than E and F (which are not revalued) following Stage 2. It also predicts that AB will receive a similar rating to EF. According to unmodified R-W, the rating of AB falls during Stage 2 as A is extinguished and B remains unaffected. Modified R-W, on the other hand, states that as V_A falls (asymptoting at 0), V_B will increase (asymptoting at λ). Given these opposing changes in associative strength, the overall associative strength of the AB compound (given by V_A+V_B) should remain roughly constant.

Note, however, that modified R-W is still an elemental theory. As such it is still constrained to predict that compounds AB and CD will receive the same rating as the elementally equivalent compounds BC and AD.

The other trial types listed in the experimental design are relevant to a different issue in associative learning theory: they are not discussed here. We were careful to ensure equal numbers of positive and negative trial types in each stage. Following Dickinson & Burke (1996), we also made sure that each subject encountered a large number of different trial types (16 in Stage 1, 8 in Stage 2). This creates a large memory load, hopefully preventing subjects from basing their ratings on inferences made from explicit episodic memories of the various trial types. Instead subjects should have to rely on associative processes to provide an "automatic" measure of causal efficacy for each cue. Using a large number of trial types makes us more confident that it is indeed associative, rather than cognitive, processes being tapped in our study.

Method

Participants Sixteen members of Cambridge University (9 female, 7 male; age 19-49) took part in the experiment.

Procedure At the start of the experiment each subject was given a sheet of instructions presenting the "allergy prediction" cover story for the experiment. They were told that in the first block they would be arrange for Mr. X to eat different meals on each day, and would monitor whether he had an allergic reaction or not as a result. In relation to the exposure trials (that do not bear on the issue at hand in this paper), subjects were told that occasionally the results of eating the foods had been lost. On these trials they would know the foods eaten in the meal, but not the result of eating those foods. They were also told that at the end of the experiment they would be asked to rate each of the foods according to how strongly it predicted allergic reactions. The 24 foods used were randomly assigned to the letters A to X in the experimental design for each subject.

On each conditioning trial, the words "Meal [meal number] contains the following foods:" followed by the two foods appeared on the screen. Subjects were then asked to predict whether or not eating the foods would cause Mr. X to have an allergic reaction, using the "x" and "." keys (counterbalanced). The screen then cleared, and immediate feedback was provided. On positive trials the message "ALLERGIC REACTION!" appeared on the screen; on negative trials the message "No Reaction" appeared. If an incorrect prediction was made, the computer beeped. On the exposure trials of Stage 2, the same message appeared, but now subjects were cued to enter the initial two letters of each of the foods. This was to ensure that they paid attention to the pairings of foods when no allergy prediction was required.

There were 16 trial types in Stage 1, and 8 in Stage 2. The order of trials was randomised over each set of 16 or 8. Participants saw each meal 8 times in Stage 1 and Stage 2. The order of presentation on the screen (first/second) within each compound pair was also randomised.

In the final rating stage subjects were asked to rate their opinions of the effect of eating a number of meals containing either one food or two on a scale from -10 to +10 (in fact, subjects were also given a short rating test at the end of Stage 1 – again the results of that test do not bear on the work presented here and so will be ignored.). They were to use +10 if the meal was very likely to cause an allergic reaction in Mr. X, -10 if eating the meal was very likely to prevent the occurrence of allergic reactions which other foods were capable of causing, and 0 if eating the meal had no effect on Mr. X (i.e. it neither caused nor prevented allergic reactions). For clarification, participants also had access to a card on which the instructions on how to use the rating scale were printed. Once a meal had been rated it disappeared from the screen and the next appeared, so that participants could not revise their opinions upon seeing later meals.

Results and Discussion

Figure 1 shows the mean rating of the casual efficacy of each of the meals of interest as judged on test. In this figure, and in the following analysis, the ratings of equivalent cues

(i.e. cues or compounds that have received an identical training history) have been averaged for each subject. Thus we averaged the ratings of AB and CD, BC and AD, A and C, B and D, and E and F. No significant differences existed between equivalent cues [$F_{max}(1,15)=1.97$, $p>0.1$].

A one-way, repeated measures ANOVA was carried out on these ratings as a preliminary to assessing the effects of interest by means of planned comparisons. There was a significant main effect of meal [$F(10,150)=13.98$, $p<0.001$].

In common with earlier studies of retrospective revaluation, we see that B and D are rated higher than E and F on test [$F(1,15)=8.72$, $p<0.01$]. Given that B, D, E and F all received exactly the same number of pairings with the outcome in Stage 1, this finding implies that the ratings of B and D have changed as a result of Stage 2 A- and C- trials. This demonstration of learning about absent cues violates the assumption of the original R-W model that learning can only proceed to cues presented on a trial. However, it is consistent with modified R-W, in which absent-but-expected cues engage the learning process with a negative sign.

The ratings for AB/CD are actually very similar to those received by EF: the difference between them is not significant [$F<1$]. This again disagrees with original R-W, which states that extinction of A should reduce the causal efficacy of the AB compound relative to EF. And again it is consistent with modified R-W, which proposes that as V_A falls on A- trials, V_B rises, such that the rating for AB will remain roughly constant. In further support of this idea we see that the ratings given to B/D do not differ significantly from those given to compounds AB/CD [$F<1$]. From the standpoint of an elemental theory, this finding implies that the associative strength of compound AB is almost entirely due to the strength of cue B, which is the prediction made by modified R-W.

Of most interest, though, is the finding that the ratings for AB/CD (the compounds actually experienced during training) are higher than those for AD/BC, even though all of these compounds are elementally equivalent. This is confirmed statistically [$F(1,15)=10.72$, $p<0.005$]. This finding is troublesome for any theory that proposes that stimuli in an associative network are represented in a wholly elemental manner, as such theories are constrained to predict that AB, CD, BC and AD will receive equal ratings on test.

Thus it is clearly insufficient to view a compound AB as simply being composed of separable A and B elements which gain associative strength independently. Instead it seems that the fact that A and B have been seen together before is important when determining the response to the AB compound, i.e. there is importance attached to the unique *configuration* of A and B cues. This heightened responding to previously experienced configurations will not apply to the BC compound, as B and C elements have never been experienced in configuration during training. Hence if configurations of cues are taken into account we can explain the finding that AB/CD are rated higher than BC/AD.

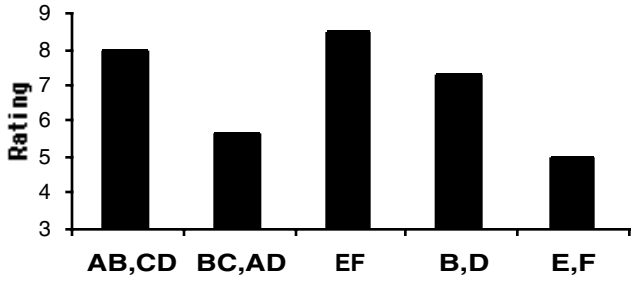


Figure 1. Mean ratings given to the cues of interest.

It is possible to modify R-W even further in order to encompass this importance of specific configurations of cues. Wagner & Rescorla (1972) suggested that, in addition to activating individual cue elements, presenting a compound stimulus also activates a unique element representing that configuration of cues (a "configural element"). Thus presentation of AB will activate elements for A and B, and also an AB element. As regards the learning process, all elements are treated in exactly the same way. If we assume that all elements have equal salience, then following AB+ trials, $V_A=V_B=V_{AB}=1/3\lambda$. Note that this is still very much an elemental theory in nature: the associative strength of a compound is given by summing the individual strengths of all of its elements, whether those elements represent compounds or single cues.

We can combine this with the idea of negative α. On A- trials following AB+ training, elements for both B and AB will have negative α: neither is present, but both are retrieved via within-compound associations. As A extinguishes (until $V_A=0$), B and AB will become more excitatory, such that the overall strength of the AB compound (given by $V_A+V_B+V_{AB}$) remains roughly constant across A- trials. There is no configural element for the BC compound, however, as this configuration has never been experienced during training. Thus the associative strength of the BC compound will be given by (V_B+V_C). Given that this compound does not receive the extra excitatory influence of a configural element, it is bound to receive a lower rating than AB. This "configural element" adaptation of Van Hamme & Wasserman's modified R-W therefore allows us to explain the finding that AB is rated higher than BC.

However, this "double modification" leads to further incorrect predictions. Firstly, it predicts that AB will be rated higher than B. Presentation of AB activates A, AB and B units. The latter two have excitatory connections to the US, and their influence will sum. Presentation of B only activates the B unit, so the excitatory influence will be less. In fact, the ratings for AB and B do not differ. Secondly, it predicts that B will receive a similar rating to BC. Given that $V_C=0$, both will rely on the B–US association for all their excitatory strength. In fact, B/D is rated significantly higher than BC/AD [$F(1,15)=5.64$, $p<0.05$].

Perhaps a consideration of context will help. Presenting USs in a context makes the context itself a weak excitor of the US. In terms of our experiment, subjects come to realise that the patient is quite prone to allergic reactions regardless of which particular foods he has eaten. Cues presented on nonreinforced trials (e.g. A and C) will become weak inhibitors of the US to counter this general excitatory influence. A and C do in fact receive negative ratings on test (mean –1.6), adding weight to this argument. If V_A and V_C are negative then we can resolve the problems outlined above. The predicted rating for AB will fall due to A's inhibitory influence: B's rating will not be affected in this way: AB and B will

receive similar ratings. Furthermore, BC will now be rated lower than B due to C's negative effect.

In summary, it would seem that an elemental rule (a) employing configural elements, and (b) allowing negative learning about absent-but-expected elements, might explain our empirical data, as long as the role of context is also taken into consideration. Figure 2 (black bars) presents simulation results for this experiment using just such a model. Comparison with the empirical data in Figure 1 shows close agreement (r^2=0.98). The exact parameters used are relatively unimportant – we simply note that it is possible for such a model to explain the patterns present in our data.

Another approach to explaining our data is to reconsider the contention of elemental theories that a stimulus compound is composed of separable A and B elements. Recently this idea has been challenged, notably by Pearce (1987), whose configural theory proposes that a compound stimulus is best viewed as a unitary event that is separate from its elements, but able to generalise to them. In other words, it would be a single, "AB" configuration that developed an associative connection to the outcome. Generalised responding to other stimuli occurs to the extent that these stimuli are similar to previously experienced configurations.

Specifically, Pearce's (1987) configural theory states that:

$$\Delta V_X = \beta_{US}(\lambda - V_X) \qquad (2)$$

where V_X (the associative strength of configuration X) is given by the sum of the *conditioned* responding to configuration X and the *generalised* responding to X as a result of its similarity to other trained configurations. The extent to which generalisation occurs between two configurations depends on their similarity:

$$S_{X,Y} = \frac{nc_X}{nt_X} \cdot \frac{nc_Y}{nt_y} \qquad (3)$$

Thus the similarity (S) between configurations X and Y is equal to the proportion of the total elements in configuration X that are common to the two configurations, multiplied by the proportion of the total elements in configuration Y that are common to the two configurations. Then:

$$\begin{matrix}\text{Generalised strength} \\ \text{from Y to X}\end{matrix} = S_{X,Y} \times V_Y \qquad (4)$$

Consider, for instance, compounds AB and BC in our experimental design. Each configuration has two elements, one of which (B) is common. Hence they will have a similarity of 0.25, so any conditioned responding to AB will generalise by a factor of 0.25 to BC.

In our Stage 1, a representation of AB will develop an associative strength of λ (again ignoring β, which will be equal for all configurations). In Stage 2 A is extinguished. A has a similarity of 0.5 to AB, and hence receives generalised strength of 0.5λ from it. In order to counteract this excitatory influence A must itself take on a strength of -0.5λ (to prevent conditioned responding when it is inappropriate).

On test, responding to AB is given by the sum of its own conditioned strength and its generalised strength from A (to which it has a similarity of 0.5). Hence:

$$V_{AB} = \lambda + 0.5(-0.5\lambda) = 0.75\lambda$$

The same holds true for CD. How about responding to AD? This configuration has never been seen before, and hence will receive only generalised strength. It has a similarity of 0.25 to AB and to CD, and a similarity of 0.5 to A. Thus:

$$V_{AD} = 0.25(\lambda) + 0.25(\lambda) + 0.5(-0.5\lambda) = 0.25\lambda$$

Hence this configural theory predicts that responding to AB will be greater than for AD.

This kind of configural theory thus seems tailor-made to explain the different ratings given to elementally equivalent old and new compounds. The problem for a configural theory such as Pearce's is that it cannot explain the occurrence of retrospective revaluation. In common with the original R-W model, Pearce predicts that B's rating should not change as a result of A- trials. Following Stage 1 AB+ trials, V_B should be 0.5λ (as B has a similarity of 0.5 to configuration AB, which will have developed an associative strength of λ). However, given that the compound AB is not seen again, its conditioned associative strength will not change, and so B's rating (which depends on generalisation of excitatory strength from the trained AB compound) will remain unchanged. Is it possible to modify a configural theory such as Pearce's to also explain the phenomenon of retrospective revaluation? The answer at present seems to be no, and we leave it for others to challenge this conclusion.

The problem for such configural rules seems to lie in their use of fixed, non-adaptive generalisation coefficients. The generalisation between two similar stimuli takes a set value that cannot change whether the two stimuli are reinforced or not. An alternative possibility is to use adaptive generalisation coefficients that vary dynamically, such that the generalisation between two similar stimuli can change on a trial-by-trial basis according to whether the two stimuli predict the same or different outcomes. For more on the value of adaptive generalisation coefficients, see McLaren (1993, 1994) and Le Pelley & McLaren (in press).

Consider the AB+, A- contingency used in our experiment. On Stage 1 AB+ trials, subjects learn that when A and B are presented together, the outcome is expected. Following these trials, they have no reason to believe that what holds for A in the presence of B should not hold true for A alone. Hence generalisation to other compounds containing A might be set high as a default. Stage 2 A- trials provide evidence against this idea, though. As a result generalisation between A- and AB+ should be reduced, in order to prevent new learning (that A alone is not reinforced) from interfering with old (that A and B in compound are reinforced). This leaves AB as a good predictor of the outcome, while simultaneously allowing complete extinction of A. The fact that the generalisation between the AB compound and its ele-

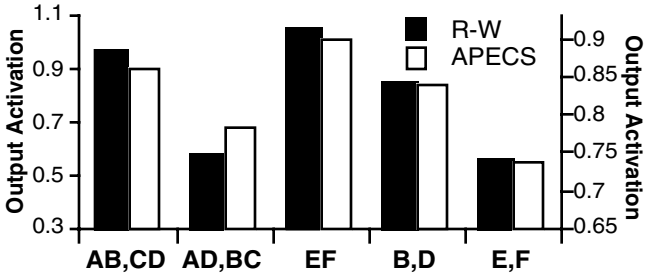


Figure 2. Simulation results for modified R-W with configural elements (black bars) and APECS (white bars).

ments is allowed to change as subjects move from Stage 1 AB+ trials to Stage 2 A- trials provides a mechanism for change in the causal efficacy of B as a result of these A- trials. As it turns out, a model employing configural representation with adaptive generalisation coefficients is indeed well-equipped to explain our empirical data.

A suitable candidate is the APECS model presented by Le Pelley & McLaren (in press). In this instantiation of APECS, each different pattern of stimulation is represented by its own hidden unit, which can equally well be termed "configural units". The mechanics of learning in APECS are similar to those of standard backpropagation (Rumelhart, Hinton & Williams, 1986), but differ in that APECS employs adaptive generalisation coefficients: once the weights appropriate to a mapping have developed, the learning in those weights can be protected against interference. This is achieved by reducing the learning rate parameter for the configural unit carrying the mapping. The effect is to "freeze" the weights to and from a certain configural unit at the value they hold immediately following experience of that configuration. Crucially, this freezing of weights to and from a certain configural unit occurs only if that configural unit has a negative error value, i.e. *if it is part of a mapping that predicts an incorrect outcome for the current input*. Specifically, APECS has different learning rate parameters for input–hidden and bias–hidden connections. The former are frozen to prevent interference; the latter remain high. Hence extinction (suppression of inappropriate responses) is achieved by an increase in the negative bias on the hidden unit carrying the inappropriate mapping, rather than by reduction of weights (which would cause the original mapping to be lost from the network). Given appropriate input cues, the negative bias on the hidden unit can be overcome and the original mapping retrieved.

Consider an AB+, A- contingency. During Stage 1, the network will learn associations from A and B input units to a hidden unit representing the AB configuration. It also learns an excitatory association from this hidden unit to the output: A and B in compound come to predict the outcome.

Now consider the inter-trial interval (ITI) between AB+ trials, when no inputs are presented. According to the logistic activation function employed with APECS, when no inputs are presented the hidden units will have an activation of 0.5 (see Rumelhart et al., 1986). This activation will feed along the AB+ pathway learnt on the preceding trial, and activate the US unit. This is obviously inappropriate when no inputs are presented. The US unit will take on a negative error, which is propagated back to the AB configural unit. As explained earlier, a negative error means that the weights to and from the hidden unit are frozen. In order to suppress the expression of the US during the ITI, the AB configural unit will therefore take on a negative bias.

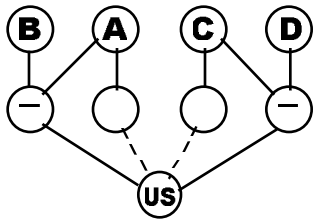


Figure 3. Associations developed by APECS following AB+, CD+ then A-, C- training. Excitatory connections are shown by solid lines, inhibitory associations by dotted lines. Negative bias on hidden units is indicated by a minus sign.

In Stage 2 the network experiences A- trials. Given that this configuration has not been seen before, a new hidden unit is recruited to carry the mapping. Of course, as a result of the associations built up during Stage 1, A has an excitatory connection to the US (via the AB configural unit). However, the US is not presented on these trials: the AB unit carries an inappropriate mapping, and so will take on a negative error. As a result its weights are frozen, and it will take on an increased negative bias in order to suppress expression of the US (i.e. to allow extinction of A). Thus the learning about the mapping from A and B to the output has not been lost from the network, it has simply become harder to retrieve. In addition, an inhibitory mapping will develop from the new A- hidden unit to the outcome in order to counter the positive activation flowing via the AB configural unit. The situation for cues A, B, C and D following Stage 2 training is shown in Figure 3.

Note that the negative bias taken on by the AB configural unit is a result of presentation of A alone leading to inappropriate output activation on A- trials. Now on test both A and B are presented together. Presentation of both cues (each with an excitatory connection to the AB configural unit) will be sufficient to overcome the negative bias built up by this unit, and so the mapping from the configural unit to the US will be expressed as before. In other words, the presence of the extra retrieval cue on test (B, compared to A alone in Stage 2) allows retrieval of the original AB+ mapping. Adaptive generalisation protects the AB+ mapping from the effect of extinction of A. Hence APECS predicts that extinction of A will have little effect on the rating received by AB, and thus that AB and EF will receive similar ratings (as is seen in Figure 1).

What if the compound BC is presented on test? As a result of the processes previously mentioned, C will be completely extinguished and so will not cause any activation of the output. Presentation of B will send some positive activation to the AB configural unit. However, without the additional positive influence of A, B alone will be unable to completely overcome the negative bias on this unit. As a result less positive activation will flow to the US than if A were also present. Thus APECS correctly predicts that AB will receive a higher rating than BC.

How about unovershadowing? Again, APECS explains the phenomenon as being a result of the attempt to minimise interference between old and new learning through adaptive generalisation. We saw that the AB configural unit starts Stage 2 with a reasonable negative bias (built up during ITIs following AB+ trials in Stage 1). This unit then takes on additional negative bias on the initial A- trials of Stage 2. However, if this negative bias is allowed to grow too much then the network will lose the information that B has in the past predicted the US, as presentation of B will be insufficient to impact on this negative bias. This would be an undesirable consequence of learning about A. In order to protect this learning, over the course of Stage 2 A- trials, as the inhibitory connection via the A-only hidden unit becomes stronger, the bias on the AB configural unit lessens.

Thus on initial A- trials, the network achieves extinction of A by suppressing the original excitatory pathway. This makes sense: given the limited evidence for the causal efficacy of A, its failure to predict the US may be a freak occurrence. It is undesirable to lose the information that A predicts the US on the basis of this limited evidence. Extinction by suppression of a pathway allows for rapid reactivation of that pathway should A now come to predict the US again. But with increasing evidence that A genuinely does not predict the outcome, the balance shifts. The original suppression is lifted to prevent loss of information about the other cues that A was trained with, which probably were the cause of the outcome originally. Extinction of A is now achieved more permanently by development of an inhibitory association to the outcome. This is sufficient to balance the increased excitation flowing from A to the US via the now less suppressed AB unit. This lesser suppression of the AB unit, meanwhile, reduces its negative bias to levels below that developed in Stage 1, meaning that presentation of B will now cause greater US activation than E or F (as the EF unit has not undergone this de-suppression). Unovershadowing is the result. For a more detailed discussion of APECS and unovershadowing, see Le Pelley & McLaren (2001).

There is a problem, however. As things stand APECS incorrectly predicts that BC should receive a rating similar to B. We can overcome this problem by considering context, as described earlier, so that A and C become weakly inhibitory. Figure 2 (white bars) shows the results of a simulation of this experiment using APECS. Again, comparing this to Figure 1 reveals close agreement between empirical and simulated data (r^2=0.98). The simulated results are actually the average of 16 simulations run with APECS, each representing a different subject. Each trial involved 1000 learning cycles. A hidden unit is defined as being "active" when it receives positive activation from the input layer. Thus if cue A is presented to the network, any hidden unit representing a configuration that includes cue A will be active. Activity extends into the period immediately following each trial, when no inputs are presented (again for 1000 learning cycles). The learning rate parameters for input–hidden and hidden–output units are both 0.8 when a hidden unit is active and has a positive error, and 0 when it is not. The parameter for bias–hidden changes is 0.3 when a hidden unit is active, 0 when it is not. Thus we make the reasonable assumption that changes due to learning take place faster than changes in memory, i.e. learning represents rapid acquisition, and memory represents a more gradual decline in retrievability. We also included an input unit representing context, that was active on every trial. Context will have a far lower salience than the foods used on each trial: we use a parameter of 0.028 for changes in weights from the context unit. The simulation results are robust under quite large variations in the parameters used.

Conclusion

Retrospective revaluation has typically been assumed to be best explained in terms of changes in the associative strengths of separable stimulus elements. Perhaps unsurprisingly, then, the most influential theories attempting to account for retrospective revaluation (e.g. Van Hamme & Wasserman's [1994] modification of R-W; Dickinson & Burke's [1996] modification of Wagner's [1981] SOP model) have been elemental in nature. The results presented here, however, suggest that this simple elemental view of retrospective revaluation is incorrect. Our data conflict with the fundamental assumption of simple elemental theories that a compound AB is best represented as being composed of separable A and B elements that gain strength independently. Instead there is a configural component involved in retrospective revaluation that is ignored in these earlier theories. It is possible to account for the data using an elemental theory modified to give a role to "configural elements". Alternatively, a model employing configural representation with adaptive generalisation also provides a good account of our results.

References

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *Quarterly Journal of Experimental Psychology, 49B*, 60-80.

Le Pelley, M.E., Cutler, D.L., & McLaren, I.P.L (2000). Retrospective effects in human causality judgment. *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (pp. 782-787). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Le Pelley, M.E., & McLaren, I.P.L. (2001). Retrospective revaluation in humans: Learning or memory? *Quarterly Journal of Experimental Psychology,* accepted subject to revision.

Markman, A. B. (1989). LMS rules and the inverse base-rate effect: Comment on Gluck and Bower (1988). *Journal of Experimental Psychology: General, 118*, 417-421.

McLaren, I. P. L. (1994). Representation development in associative systems. In J.A. Hogan & J.J. Bolhuis (Eds.), *Causal mechanisms of behavioural development* (pp. 377-402). Cambridge: Cambridge University Press.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review, 94*, (61-73).

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A.H. Black & W.F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart, J.L. McLelland & the PDP Research Group (Eds.), *Parallel Distributed Processing* (Vol. 1, pp. 318-362). Cambridge, MA: MIT Press.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation, 25*, 127-151.

Wagner, A.R., & Rescorla, R.A. (1972). Inhibition in Pavlovian conditioning: Application of a theory. In R.A. Boakes & M.S. Halliday (Eds.), *Inhibition and Learning (pp. 301-336).* New York: Academic Press.

Wagner, A. R. (1981). SOP: A model of automatic memory processing in animal behaviour. In N.E. Spear & R.R. Miller (Eds.), *Information processing in animals: Memory mechanisms* (pp. 5-47). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Costs of Switching Perspectives in Route and Survey Descriptions

Paul U. Lee (pauly@psych.stanford.edu)
Department of Psychology, Bldg. 420
Stanford, CA 94305 USA

Barbara Tversky (bt@psych.stanford.edu)
Department of Psychology, Bldg. 420
Stanford, CA 94305 USA

Abstract

Two experiments examined perspective switching in comprehension and retrieval of spatial information. Participants read route or survey descriptions of environments line by line. Reading times were recorded. For half of the descriptions, the perspective of the last sentence was switched. True/false verification of sentences from both perspectives followed the descriptions. Switching perspective increased reading times but increased verification times only for survey sentences. This suggests that perspective switching exacts a cost in comprehension, but that the cost dissipates after information retrieval, especially for route descriptions. The second study examined which aspects of perspective, viewpoint or terms of reference, accounted for switching costs by using hybrid descriptions. Switching terms of reference slowed reading times more than switching viewpoint. Together, the experiments suggest that switching perspective plays a role in comprehension that diminishes with repeated retrieval. They also point to a fundamental asymmetry between route and survey perspectives, one that depends on orientation.

Introduction

All animals, human or otherwise, must be able to gain knowledge about their surroundings in order to survive. Like other animals, humans often gain this knowledge first-hand by navigating through their immediate surroundings. However, unlike other animals, humans also have the ability to transmit this knowledge to others by language.

Whether the world is experienced first-hand or through descriptions, the world is viewed from a particular perspective. In text comprehension, maintaining a consistent perspective makes the text more coherent and comprehensible to the readers (Black, Turner, & Bower, 1979).

Although there is a general agreement that we perceive the world from specific viewpoints, the role of perspectives on spatial memory is less certain. In particular, there is a considerable debate about whether or when spatial relations are encoded independently of perspective.

If spatial memory is formed in a perspective-dependent manner, it should be more accessible from one perspective over the others. Therefore, when people describe spatial layouts from memory, they should prefer to maintain a consistent perspective in their descriptions since memory associated with a specific perspective is more accessible than others (Levelt, 1982). There is some evidence that supports this view. When readers take a particular viewpoint during text comprehension, they later remember the information better from that viewpoint (Black, Turner, and Bower, 1979; Abelson, 1975).

Other evidence suggests that for constrained well-learned environments, spatial memory can equally be accessible from multiple perspectives. Taylor and Tversky (1992) demonstrated that when people learned route or survey descriptions of spatial layouts and later were asked inference questions about them, they were as fast and accurate to questions from the read perspective as from the new perspective. In spontaneous descriptions of naturalistic environments, people mixed route and survey perspectives about half the time (Taylor & Tversky, 1996). The choice of a perspective also seems to depend on the pragmatics of the situation much more than an inherent bias toward any particular perspective. During a conversation, speakers often use not only their own perspective, but also the perspective of their addressee's or some perspective independent of both (Schober 1993; Tversky, Lee, & Mainwaring, 1999).

In visual cognition, there is a similar inquiry about the nature of spatial memory derived from perception and navigation. There is some evidence that spatial relations are encoded in a viewpoint-dependent manner (Diwadkar & McNamara, 1997; Shelton & McNamara; Rieser, 1989). For example, Diwadkar and McNamara (1997) had participants study the locations of objects in a room from a single perspective and then learn to recognize the layout from three other views. A recognition test of the layout from different viewpoints showed faster response times for the learned viewpoints than the novel viewpoints. However, memories of large-scale spaces seem to encode spatial relations in a viewpoint-independent manner, in which familiar and

novel views of spatial layouts are equally accessible (Evans & Pezdek, 1980; Presson, DeLange, & Hazelrigg, 1989).

So far, the evidence seems to suggest that maintaining a consistent perspective is important when learning a new environment, but the effect of perspective is less clear for retrieving well-learned environments. The present study is an attempt to understand the role of perspectives in spatial descriptions during both on-line comprehension and subsequent retrieval of the layouts from memory.

As an extension to Taylor and Tversky study (1992), we studied acquisition of environments by text that maintained a consistent perspective or switched perspectives. Testing was from same or switched perspective. We expected that the cost of switching perspective be large when under construction of spatial mental models but diminished after repeated retrieval.

Experiment 1

Method

Subjects. Thirty-nine undergraduates, 18 male and 21 female, from Stanford University participated individually in partial fulfillment of a course requirement. The criterion of 67% correct response eliminated the data of three men and four women.

Materials. Descriptive texts were prepared for sixteen fictitious environments. Each environment consisted of two intersecting roads and three adjacent landmarks. The descriptions were given either in a route or a survey perspective.

Route Description

Go east on High St and you will intersect with a much narrower Green Ave.
Turn right on Green Ave and on your right, you will see the stock market.
Past the stock market, on your right on Green Ave, you will see the mortgage bank.
On your right on Green Ave, past the mortgage bank is the legal firm.

Survey Description

High St runs east-west, intersecting a much narrower Green Ave, which runs north-south.
South of High St on the west side of Green Ave is the stock market.
South of the stock market on the west side of Green Ave is the mortgage bank.
On the west side of Green Ave, south of the mortgage bank is the legal firm.

Figure 1: Route and Survey Descriptions

A route perspective takes an imagined navigator through an environment describing landmarks relative to the navigator in terms of *left* and *right*. A survey perspective takes bird's eye view of the environment describing landmarks relative to each other in terms of cardinal directions.

Each description consisted of two introduction sentences, followed by four sentences that described the spatial layout of the environment. Figure 1 shows examples of the spatial descriptions.

In order to examine the perspective switching cost during on-line comprehension, the last sentence of the study phase was presented either in the same perspective as the preceding descriptions or in a new perspective. Figure 2 shows the perspective switch in both directions for the target sentence.

Perspective Switch: Route to Survey

Past the stock market, on your right on Green Ave, you will see the mortgage bank.
On the west side of Green Ave, south of the mortgage bank is the legal firm.

Perspective Switch: Survey to Route

South of the stock market on the west side of Green Ave is the mortgage bank.
On your right on Green Ave, past the mortgage bank is the legal firm.

Figure 2: Perspective Switch during On-line Comprehension

Four statements (i.e. two statements each for route and survey perspective) followed the target sentences for true/false verification. These questions provided assurance that participants formed accurate mental models of the spatial layouts. Since the questions used in both perspectives, half required perspective switching with respect to the study perspective. All of the questions were inference questions, querying the spatial relations that could be inferred but were not directly specified in the descriptions. An example of route inference statement is "The stock market is on your right when you face the mortgage bank from Green Ave.", and an example of a urvey statement is "The mortgage bank is north of the legal firm and west of Green Ave.".

Design and Procedure. Subjects were told that they would read descriptions of various environments. They were asked to study and remember them because after each scene, they will be given true/false questions to test their memory of the scene. They were then given a practice trial, so that they would be familiar with the overall nature of the experiment. The trial consisted of a route and a survey environment followed by four test questions.

For the actual trial, subjects read sixteen texts, i.e. eight route and eight survey descriptions, and then answered four true/false questions for each description. The order of presentation and the assignment of experimental condition to environments were randomized across subjects. The texts appeared on the screen one sentence at a time. Each sentence remained on the screen until participants pressed a key to indicate that they were ready to move on to the next sentence. The reading time for each sentence was recorded. After reading a description, participants answered four test questions by pressing assigned keys for true and false. Both

response time and accuracy were recorded for each question. The question order was randomized across subjects. The experiment was conducted on a Apple PowerMac computer controlled by PsyScope software package (J.D. Cohen, MacWhinney, Flatt, & Provost, 1993).

Results

The reading time during the study phase indicates the amount of time that subjects needed to comprehend the descriptions. Using repeated measures design, reading times (RT) per syllable were compared between route and survey perspective for all study sentences except for target sentences. Subjects studied survey texts longer (351 msec/syllable) than route texts (311 msec/syllable). $F(1, 31) = 12.39$, $p < 0.001$.

The target sentence was the last sentence in the study phase. Half of the target sentences switched perspective from preceding study descriptions and the other half kept the same perspective. Target sentences were analyzed for two factors: perspective of the target sentence and perspective consistency.

On the average, subjects read route targets marginally faster (456 msec/syllable) than survey targets (513 msec/syllable; $F(1,31) = 3.02$, $p < 0.10$). They read targets that kept the same perspective much faster (397 msec/syllable) than the targets with a new perspective (572 msec/syllable; $F(1,31) = 30.89$, $p < 0.00001$.

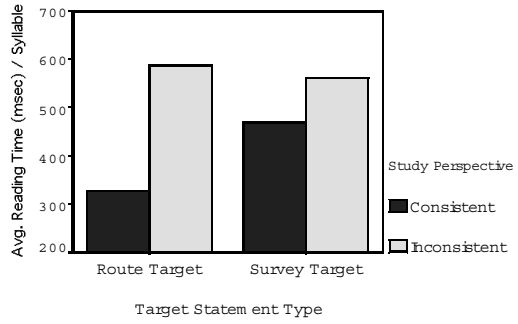


Figure 3: RT/syllable for Target Sentences: Effects of Perspective Switch during On-line Comprehension

Furthermore, there was an interaction between target perspective and perspective consistency ($F(1,31) = 10.43$, $p < 0.003$). Figure 3 illustrates that on the average, reading a route target sentence that was preceded by a description with a consistent route perspective was much faster (0.33 s/syl) than reading a route target sentence preceded by a survey description (0.59 s/syl). Although a same general pattern holds for survey target sentences, the difference in RTs is smaller (0.47 s/syl and 0.56 s/syl for consistent and inconsistent perspectives respectively).

For the true/false verification statements, we collected both response time (RT) and accuracy data. We analyzed response times for question perspective and study perspective. We also checked for effects of target perspective since it differed from study perspective half the trials. These analyses are reported for correct RTs to true questions, since RT analyses for incorrect answers or false questions yielded no significant results. It seems that subjects needed to correctly verify an accurate mental model to produce a consistent perspective effect.

Route questions were verified faster and more accurately than the survey questions (423 ms/syl for route, 476 ms/syl for survey; $F(1,31) = 9.02$, $p < 0.005$; 83% for route, 75% for survey, $F(1,31) = 7.90$, $p < 0.008$). There was no effect of target perspective on the response time ($F(1,31) = 0.0005$; $p > 0.98$).

The effect of perspective consistency was also significant but this was due only to survey questions, as there was a significant interaction between the question perspective and its consistency with the study perspective. $F(1,31) = 9.71$, $p < 0.004$. Subjects responded equally fast to route statements, regardless of perspective (426 ms/syl for route vs. 419 ms/syl for survey) but were faster to survey questions when they had studied from survey perspective (429 ms/syl; 523 ms/syl for route) (see Figure 6). Accuracy data for this interaction was not significant but was consistent with the RT results, assuring us that there is no speed-accuracy tradeoff.

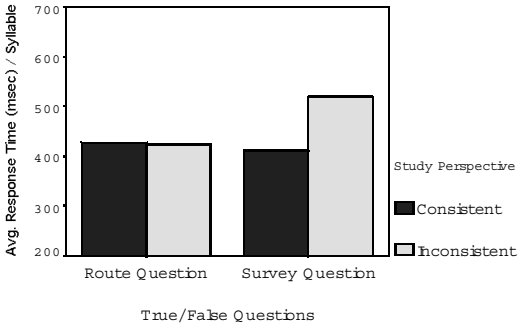


Figure 4: RT/syllable for T/F Questions: Effects of Perspective Switch during Memory Retrieval

Discussion

As predicted, switching perspective during on-line text comprehension resulted in a longer reading time, supporting cognitive costs for perspective switching during acquisition.

The interaction between target type and perspective consistency suggests that perspective consistency is more important for route descriptions. Since a route perspective is relative to a moving referent, the knowledge of current location and orientation of the

referent is crucial to processing the incoming route information. Switching from survey to route descriptions forces the reader to establish the current orientation of the referent without the benefit of inferring this knowledge from the previous text. Since survey descriptions are based on a fixed orientation perspective, consistency has no advantage over switched perspective in establishing the orientation information.

For the true/false verification statements, there was an effect of perspective switch, but the effect size was diminished from that of on-line comprehension. Furthermore, the effect completely disappeared for the route statements. The results indicate that retrieval from spatial memory is less perspective-dependent than on-line comprehension. The diminished effects of perspective for memory together with previous findings of perspective-independent memory (Taylor & Tversky, 1992) suggest that conversion from perspective-dependent to perspective-independent spatial memory may be a gradual process. If so, the present results indicate that route and survey perspective have different timetable for this conversion.

Finally, subjects took longer to study survey descriptions than route descriptions. This result is opposite that of Taylor and Tversky (1992). Although readers of route perspectives may have had greater cognitive loads due to continual updates and integration of changing location and orientation of the referent, they still read the description faster. It may be that any such difficulty is compensated by other factors, such as greater familiarity with intrinsic spatial terms (e.g. left/right) than extrinsic terms (e.g. north/south) in everyday route descriptions.

Experiment 2

We have established that switching between route and survey perspectives during study has cognitive costs. The nature of these costs seems different for route and survey perspectives. What are some of the factors that might account for these differences?

Route and survey descriptions differ in at least two ways. One is the way orientation is described. Route descriptions use intrinsic spatial terms, such as *left* and *right,* which change with the changing orientation of the navigator. Thus, they adopt a person-centered reference frame. Survey descriptions use extrinsic spatial terms, such as *north* and *south*, which fix the orientation in space, adopting an environment-centered reference frame.

Another difference is the viewpoint of the observer. Route perspectives are embedded within the environment, whereas survey perspective are external and above the environment.

In this experiment, we created hybrid descriptions that take route-like viewpoints but update its orientation using extrinsic terms. For example, a route description such as "Go down the street, turn right, and the building will be on your right" can be converted to a hybrid description like "Go north on the street, turn east, and the building will be south of you."

Using hybrid descriptions, we can examine if the perspective switching costs are due to changes in orientation terms or changes in viewpoint. If orientation terms are crucial, we expect hybrid results to mirror the survey results. If viewpoint is important, we expect hybrid results to mirror the route results.

Method

Subjects. Sixty-four undergraduates, 30 male and 34 female, from Stanford University participated individually in partial fulfillment of a course requirement. 67% accuracy criterion eliminated two men and eight women.

Materials. The stimuli were similar to those used in Experiment 1 with few changes. In addition to route and survey descriptions, hybrid descriptions were added for the study and the test phase. Figure 5 shows an example of a hybrid description.

Hybrid Description

Go east on High St and you will intersect with a much narrower Green Ave.
Turn south on Green Ave and west of you, you will see the stock market.
Past the stock market, on the west side of Green Ave, you will see the mortgage bank.
West of you on Green Ave, south of you past the mortgage bank is the legal firm.

Figure 5: Description in Hybrid Perspective

Both descriptions and targets were given in each of three perspectives. Two fictitious environments were added to the sixteen environments in Experiment 1 to match the number of environments with the experimental conditions. In addition, four true/false questions per environment were reduced to three, one each from route, survey, and hybrid perspective. Finally, the number of practice trials was increased to three, one for each description type.

Design and Procedure. Except for the changes in the materials described in the previous section, the procedure is identical to Experiment 1.

Results

The reading time results replicated those of Experiment 1. The new results for this experiment show how switching perspectives with hybrid descriptions affect on-line comprehension and memory.

We analyzed reading times for the target sentences for two factors: perspective of the target sentence and perspective consistency with study. Each factor has three perspectives: route, hybrid, and survey.

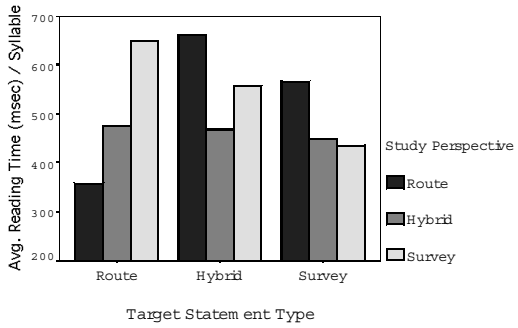


Figure 6: RT/syllable for Target Sentences: Effects of Perspective Switch during On-line Comprehension

There was an interaction between target perspective and study perspective (see Figure 6). $F(4,212) = 20.55$, $p < 2e\text{-}11$). Planned contrasts on hybrid target sentences showed faster reading times for hybrid (468 msec/syl) than route descriptions (661 msec/syl; $t(212) = 4.89$, $p < 1e\text{-}6$), and reading times for hybrid was marginally faster than survey descriptions (556 msec/syl; $t(212) = 2.16$, $p < 0.02$; Bonferroni group $p_{crit} = 0.0025$). Direct comparison between route and survey descriptions showed that a perspective switch from route to hybrid took marginally longer than a switch from survey to hybrid ($t(212) = 2.73$, $p < 0.004$; $p_{crit} = 0.0025$).

Similarly, route target sentences were read faster after route (357 msec/syl) than either hybrid (474 msec/syl; $t(212) = 3.04$, $p < 0.0015$) or survey descriptions (649 msec/syl; $t(212) = 4.55$, $p < 5e\text{-}6$). In addition, a perspective switch from survey to route took significantly longer to understand than a switch from hybrid to route ($t(212) = 7.60$, $p < 1e\text{-}11$).

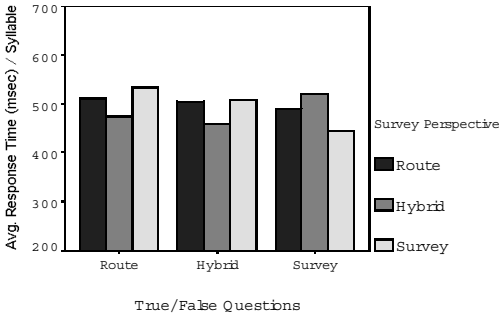


Figure 7: RT/syllable for T/F Questions: Effects of Perspective Switch during Memory Retrieval

Finally, survey target sentences were read faster when preceded by survey descriptions (435 msec/syl) than by route descriptions (566 msec/syl; $t(212) = 3.41$, $p < 0.0004$), but not by hybrid descriptions (448 msec/syl; $t(212) = 0.34$, $p > 0.35$). A direct comparison between route and hybrid descriptions showed a perspective switch from route to survey to take longer than a switch from hybrid to survey ($t(212) = 3.07$, $p < 0.0015$). Overall, hybrid statements behaved more like survey statements than like route statements.

True/false verification statements failed to replicate the results from Experiment 1, as there was no significant main or interaction effects (see Figure 7).

Discussion

The hybrid perspective provided an "intermediate" perspective between route and survey perspective. Its viewpoint was embedded like a route perspective but its reference terms were like survey perspective (i.e. north, south, east, and west). Which of these two factors would play a role in exacting cognitive costs during comprehension? The answer turned out to be both.

Reading times for target sentences were longer when the perspective was switched from route to hybrid than when consistent. Since the route and the hybrid descriptions were identical except for orientation terms, orientation terms contributed to the perspective switching costs. Similarly, when perspective was switched from survey to hybrid, longer reading times resulted. Since a survey description differs from a hybrid in the viewpoint of the observer, viewpoint also contributed to the perspective switching costs, although the effect size was much smaller than that of the orientation terms.

Target reading times for the route statements confirmed this hypothesis. Switching from the hybrid to route perspective increased the reading times, implicating orientation terms or reference frames. Since a switch from survey to route required a change in both orientation terms and the viewpoint, reading times were even longer than for a switch from hybrid to route. The advantage of hybrid over survey study perspectives for reading route targets further supported the significance of the viewpoint in the perspective switching costs.

For the survey targets, the overall results corroborated the other findings, except that there was no perspective cost when readers switched from hybrid to survey perspective. This suggests that orientation terms exact a greater cost than viewpoint.

Overall, there were strong and consistent effects of perspective switching costs due to changes in orientation terms. The effects due to viewpoint changes were weaker, since there was a significant effect for the route targets, a marginal effect for the hybrid targets, and no effects for the survey targets.

During the sentence verification, all effects between conditions disappeared. These results were consistent with Taylor and Tversky (1992) but differed slightly

from the results of Experiment 1. Both experiments are consistent with the claim that perspective effects diminish during the retrieval phase compared to the study phase, especially when the retrieval is done repeatedly from multiple perspectives. Answering true/false questions from three different perspectives might have accelerated the process.

Conclusion

When people describe a large-scale environment, they typically adopt route or survey perspectives, or a combination of both (Taylor & Tversky, 1996). These two perspectives are also readily understood, suggesting that they capture a natural way of understanding the world.

Previous work suggested that mental representations of constrained, well-learned environment acquired from descriptions are perspective-free, that is, statements about the environments from either perspective are verified equally quickly and accurately, irrespective of study perspective (Taylor & Tversky, 1992). However, other studies have suggested perspective-dependent representations (Diwadkar & McNamara, 1997).

The present experiment provided evidence that maintaining a consistent perspective during learning facilitates acquisition of new environments, that is, switching perspective exacts a cost in reading times. However, after acquisition and during testing for memory of the environment, costs of switching perspectives diminish considerably (Exp 1) or disappear (Exp 2). Perspective-independent responding could indicate that the mental representations are more abstract than any particular perspective, allowing equally efficient retrieval from either perspective. Alternatively, it could indicate multiple representations or increased efficiency of comprehending various perspectives.

Spatial perspectives include both viewpoints and terms of reference, as well as other factors, such as referent object. In particular, survey descriptions take an overhead viewpoint and use the cardinal direction terms of reference. Route descriptions take viewpoints within environments and use intrinsic direction terms, *left, right, front,* and *back.* In the second experiment, a hybrid description was constructed using a route viewpoint and survey reference terms. Switching reference terms exacted a greater cognitive cost than switching viewpoint, suggesting that overall reference frame is more critical in spatial descriptions than viewpoint.

Acknowledgments

This research was supported by Stanford Graduate Fellowship to the first author and by Office of Naval Research, Grant Number NOOO14-PP-1-O649, to the second author. We are grateful to Herb Clark and Gordon Bower for helpful discussion.

References

Abelson, R. (1975). Does a story understander need a point of view? In R. Schank & B. L. Nash-Webber (Eds.), *Theoretical issues in natural language processing.* Washington, DC: Association for Computational Linguistics.

Black, J. B., Turner, T. J., & Bower, G. H. (1979). Point of view in narrative comprehension, memory and production. *Journal of Verbal Learning and Verbal Behavior, 18*, 187-198.

Cohen, D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavior Research Methods, Instruments, & Computers. 25,* 257-271.

Diwadkar, V. A., & McNamara, T. P. (1997). Viewpoint dependence in scene recognition. *Psychological Science, 8(4),* 302-307.

Evans, G. W., & Pezdek, K. (1980). Cognitive mapping: Knowledge of real-world distance and location information. *Journal of Experimental Psychology: Human Learning & Memory, 6,* 13-24.

Levelt, W. J. M. (1982). Linearization in describing spatial networks. In S. Peters & E. Saarinen (Eds.), *Processes, beliefs, and questions.* Dordrecht: Reidel.

Presson, C. C., DeLange, N., & Hazelrigg, M. D. (1989). Orientation specificity in spatial memory: What makes a path different from a map of the path? *Journal of Experimental Psychology: Learning, Memory, & Cognition, 15,* 225-229.

Rieser, J. J. (1989). Access to knowledge of spatial structure at novel points of observation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 887-897.

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition, 47(1),* 1-24.

Shelton, A. L., & McNamara, T. P. (1997). Multiple views of spatial memory. *Psychonomic Bulletin & Review, 4(1),* 102-106.

Taylor, H. A., & Tversky, B. (1992). Spatial mental models derived from survey and route descriptions. *Journal of Memory & Language, 31(2),* 261-292.

Taylor, H. A., & Tversky, B. (1996). Perspective in spatial descriptions. *Journal of Memory & Language, 35(3),* 371-391.

Tversky, B., Lee, P., & Mainwaring, S. (1999). Why do speakers mix perspectives?.*Spatial Cognition and Computation, 1,* 399-412.

A Connectionist Investigation of Linguistic Arguments from the Poverty of the Stimulus: Learning the Unlearnable

John D. Lewis (jlewis@crl.ucsd.edu)
Department of Linguistics, McGill University; 1085 Dr. Penfield Avenue
Montreal, Quebec H3A1A7 Canada
&
Center for Research in Language, UC San Diego; 9500 Gilman Dr.
La Jolla, CA 92093-0526 USA

Jeffrey L. Elman (elman@crl.ucsd.edu)
Center for Research in Language, UC San Diego; 9500 Gilman Dr.
La Jolla, CA 92093-0526 USA

Abstract

Based on the apparent paucity of input, and the non-obvious nature of linguistic generalizations, Chomskyan linguists assume an innate body of linguistically detailed knowledge, known as Universal Grammar (UG), and attribute to it principles required to account for those "*properties of language that can reasonably be supposed not to have been learned*" (Chomsky, 1975). A definitive account of learnability is lacking, but is implicit in examples of the application of the logic. Our research demonstrates, however, that important statistical properties of the input have been overlooked, resulting in UG being credited for properties which are demonstrably learnable; in contradiction to Chomsky's celebrated argument for the innateness of structure-dependence (*e.g.* Chomsky, 1975), a simple recurrent network (Elman, 1990), given input modelled on child-directed speech, is shown to learn the structure of relative clauses, and to generalize that structure to subject position in *aux*-questions. The result demonstrates that before a property of language can *reasonably* be supposed not to have been learned, it is necessary to give greater consideration to the indirect positive evidence in the data — and that connectionism can be invaluable to linguists in that respect.

Introduction

Chomskyan linguists argue that language acquisition cannot strictly be a matter of learning — the child's target grammar is "*hopelessly underdetermined by the fragmentary evidence available*" (Chomsky, 1968) — rather it must rest on a set of innate linguistic principles; the goal of the Chomskyan linguist is to determine the contents of this set, known as Universal Grammar (UG). The idea is to attribute to UG all and only the principles required to account for those "*properties of language that can reasonably be supposed not to have been learned*" (Chomsky, 1975). Learning theory is thus of central importance to the enterprise, but, oddly, a definitive account of the notion of learning that Chomskyans adopt is lacking, and is given only implicitly in the examples of the principles attributed to UG. Statistical approaches, however, and the notions of generalization and analogy have been explicitly rejected as irrelevant (Chomsky, 1975). In this paper we demonstrate that this rejection is a serious error — that UG has been attributed with principles to account for properties of language that are demonstrably learnable from the statistical properties of the input.

Chomsky's celebrated argument for the innateness of the principle of structure-dependence (Chomsky, 1975) serves as an example. Chomsky claims that, during the course of language acquisition, children entertain only hypotheses which respect the abstract structural organization of language, though the data may also be consistent with structure-independent hypotheses, *i.e.* relationships over utterances considered only as linearly ordered word sequences. As support for this claim, Chomsky notes that though questions like (1) are apparently absent in the child's input, questions like (2) are never erroneously produced — a claim subsequently

1) *Is the man who is smoking crazy?*
2) **Is the man who smoking is crazy?*

empirically tested and substantiated by Crain and Nakayama (1987, also see Crain 1991). Chomsky suggests that it is reasonable to suppose that children derive *aux*-questions from declaratives, and exposed to only simpler structures, might hypothesize either of two sorts of rules: a structure-independent rule — *i.e.* move the first '*is*' — or the correct structure-dependent rule. Chomsky claims that "*cases that distinguish the hypotheses rarely arise; you can easily live your whole life without ever producing a relevant example to show that you are using one hypothesis rather than the other one*" (Piatelli-Palmarini, 1980). The fact that children do not produce questions like (2), despite that the correct rule is supposedly more complex, and that the learner might not encounter the relevant evidence leads Chomsky to suggest that "*the only reasonable conclusion is that* UG *contains the principle that all such rules must be structure-dependent*" (Chomsky, 1975).

As a number of researchers have noted, however, there are several weaknesses in this argument. Slobin (1991), for instance, points out that the conclusion rests on the assumption that *aux*-questions are derived from declar-

atives by movement — an assumption which lacks justification — as well as on the equally questionable assumption of the autonomy of syntax. The argument has also been widely criticized for its reliance on the extremely limited conception of learning as hypotheses generation and testing. And the premise that the relevant evidence is not available to children has repeatedly been argued to most likely be false. As Sampson (1989) points out, evidence to distinguish the two hypotheses is provided by any utterance in which *any* auxiliary precedes the main clause auxiliary; thus evidence is available not only in questions like "*Is the jug of milk that's in the fridge empty?*" (from Cowie, 1998), but also "*Is the ball you were speaking of in the box with the bowling pin?*", or "*Where's this little boy who's full of smiles?*", or even "*While you're sleeping, shall I make the breakfast?*" None of these forms seem to be of the sort that a person might go for long without encountering; the latter three examples, in fact, are taken from the CHILDES database,[1] and Pullum and Scholz (2001) estimate that such examples make up about one percent of a typical corpus.

These are strong criticisms, but a conclusive counterargument, or an alternate account of the acquisition of *aux*-questions remains to be given. This paper builds on recent work with simple recurrent networks (SRNs; Elman 1990) to close this gap — *i.e.* to provide a proof that the correct form of *aux*-questions is learnable from data uncontroversially available to children.

Figure 1 shows the general structure of an SRN. The recurrent connections from the hidden layer to the context layer provide a one-step state memory. At each time step the activation values of each of the hidden units is copied to the corresponding unit in the context layer, and the connections from the context layer back to the hidden layer make these values available as additional inputs at the next time step. The network receives its input sequentially, and at each step attempts to predict the next input. At the outset of training, the connection weights and activation values are random, but to the extent that there are sequential dependencies in the data, the network will reduce its prediction error by building abstract representations that capture these dependencies. Structured representations thus emerge over time as a means of minimizing error.

Elman (1991, 1993) provided such a network with a corpus of language-like sentences which could be either simple (transitive or intransitive), or contain multiply embedded relative clauses (in which the head noun could be either the subject or object of the subordinate clause). The input was presented as word sequences, where words were represented as orthogonal vectors — a localist representation — so that no information about either the words or the grammatical structure was supplied; thus the network had to extract all information (*e.g.* the grammatical categories, number agreement, subcategorization frames, and selectional restrictions) from regularities in the input. The network learned the structure of such sentences so as to predict the correct agreement patterns between subject nouns and their corresponding verbs, even when the two were separated by a relative clause with multiple levels of embedding, *e.g.* *boys who like the girl who Mary hates hate Mary*.[2,3]

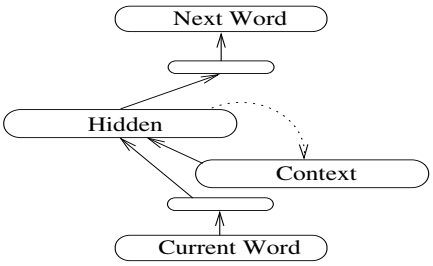


Figure 1: An SRN. Solid lines represent full connectivity; the dashed line indicates unit-to-unit connections. The unlabeled layers are reduction layers.

Such networks have also been shown to go beyond the data in interesting ways. Elman (1998) and Morris et al. (2000) showed that SRNs induce abstract grammatical categories which allow both distinctions such as *subject* and *object*, and generalizations such that words which have never occurred in one of these positions are nonetheless predicted to occur, if they share a sufficient number of abstract properties with a set of words which have occurred there.

Together these results suggest that an SRN might be able to learn the structure of relative clauses, and generalize that structure to subject position in *aux*-questions — and thus to learn the aspect of grammar in question despite not having access to the sort of evidence that has been assumed necessary. This paper reports on simulations which show that this is the case. An initial experiment verifies that the two results combine in the required way; then an SRN is shown to generalize from training sets based on CHILDES data to predict (1), but not (2). This result clearly runs counter to Chomsky's argument, and thus both draws into question the validity of poverty of the stimulus arguments in general, and shows that neural networks provide a means of assessing just how impoverished the stimulus really is.

Abstractions and Generalization

Training sets similar to those used by Elman (1991, 1993) were used to test whether an SRN would generalize to predict relative clauses in subject position in *aux*-questions from data which contained no such questions. An artificial grammar was created such that it generated *a) aux*-questions of the form 'AUX NP ADJ?',

[1]The second through fourth examples are from Brown's Adam, Korman's' St, and Manchester's Anne, respectively.

[2]The network succeeded only if either the input was structured, or the network's memory was initially limited, and developed gradually.

[3]An SRN's performance with such recursive structures has also been shown to fit well to the human data (Christiansen and Chater, 1999).

and *b)* sequences of the form 'A_i NP B_i', where A_i and B_i were of varying content and length. Proper names and NPs of the form 'DET (ADJ) N (PP)' were generated in both types, and NPs with relative clauses were generated for the 'A_i NP B_i' type, but were restricted from appearing in *aux*-questions. Some representative examples are given in Figure 2.

A_i *Mummy* B_i	*is Mummy beautiful?*
A_i *the dog* B_i	*is the dog hungry?*
A_i *the little girl* B_i	*is the little girl pretty?*
A_i *the cat on the mat* B_i	*is the cat on the mat fat?*
A_i *the boy who is smiling* B_i	*

Figure 2: Examples of the various types of utterances generated by the artificial grammar.

A three-stage training set was generated from this grammar, with the degree of complexity in NPs increasing at each stage, and the percentage of *aux*-questions decreasing — crudely approximating the structure of child-directed speech. Names constituted 80% of the NPs in the first set, and the remaining 20% was shared among the other NP forms (such that the more complex the form, the fewer the instances of it), with relative clauses making up only 1%; there were 40% *aux*-questions, and 60% 'A_i NP B_i' forms. In the second set, names constituted 70% of the NPs, relative clauses made up 2.5% of the remainder, and the percentage of *aux*-questions decreased to 30%. And in the third set, 60% of the NPs were names, relative clauses made up 5% of the remainder, and the percentage of *aux*-questions decreased to 20%. Each training set consisted of 50,000 examples. An SRN was trained on each set successively, for 10 epochs each, and tested with the structures in (1) and (2) after each epoch.[4] The network received the input in the same form as used by Elman (1991, 1993), *i.e.* a localist representation was used, and the data was presented one word at a time.

Figure 3 shows the networks predictions (after the third stage of training) for successive words of the question "*Is the boy who is smoking crazy?*" As should be expected, the network predicts an AUX as a possible first word, a name or a DET as a continuation when presented with '*is*', and a noun or an adjective as possibilities after '*is the*'. These sequences all occur in the training sets. But, following presentation of '*is the boy*', not only is an adjective or a preposition predicted, but also a relativizer — a sequence which never occurs in the training sets. And upon presentation of '*who*' the network predicts an AUX, and when given '*is*', predicts a participle; the network has thus generalized to predict the

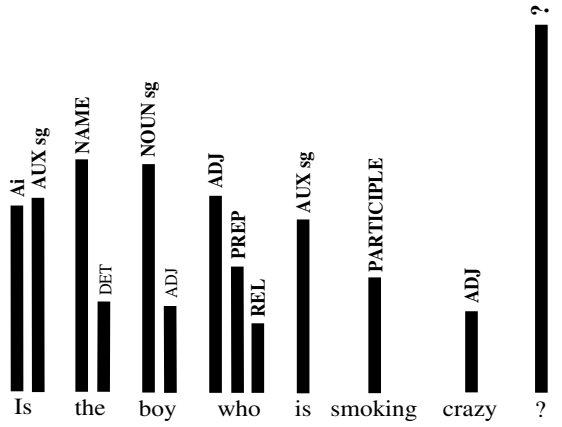


Figure 3: The SRN's categorized predictions for the test sentence "*Is the boy who is smoking crazy?*" Target words appear under the network's predictions; and the strength of the predictions is represented vertically.

relative clause.[5] The network does not, of course, make the predictions corresponding to the ungrammatical form in (2) — *i.e.* the network does not predict a participle following '*who*'; the training sets do not contain copula constructions, and so there can be no hypothesis of a movement derivation. Rather, the network has apparently formed an abstract representation of NPs which includes NPs with relative clauses. That this is so is shown by the networks prediction of an adjective when presented with '*is the boy who is smoking* ____'; the sequence '... PARTICIPLE ADJ ...' never occurs in the training sets, and thus the prediction indicates that the network has formed an abstract representation of *aux*-questions, and generalized over the NP forms.

That the data available to children are sufficient to provide for this generalization, however, remains to be shown.

Child-Directed Speech

There are a number of features of child-directed speech that run counter to the notion that the child's input is "*meager and degenerate*" (Chomsky, 1968) — *i.e.*, that appear to be important for language acquisition, and particularly for the issue at hand. Complexity increases over time — which has been shown to be a determinant of learnability (*e.g.* Elman, 1991, 1993) — and there are also arguably meaningful shifts in the distribution of types, and the limitations on forms.

The increasing complexity of the child's input is especially relevant to the problem here, since it is directly linked to the frequency of occurrence of relative clauses.

[4]The networks were simulated with *LENS* (Rohde, 1999), and trained with a fixed learning rate of 0.01, using a variation of cross entropy which assigned smaller errors for predicting incorrectly than for failure to predict. The architecture shown in Figure 1 is used, with 100 input and output units, 50 units in the reduction layers, and 500 units in both the hidden and context layers.

[5]The fact that the network predicts '*who*' given '*is the boy*' is, on its own, not enough — early in training, the network will make this prediction, but when presented with '*who*' will predict a '*?*', apparently mistaking the relativizer for an adjective. That the network *is* predicting a relative clause is shown by the fact that it predicts '*is*' when subsequently given '*who*', and a participle when then given '*is*'. Since participles are restricted to only occur in relative clauses, the latter is decisive.

Complexity in the child's input is introduced in a way akin to the staged presentation of data used to train the network in the experiment described above; Figure 4 charts the occurrences of tagged relative clauses — *i.e.* marked with '*who*' or '*that*' — found in child-directed speech in the CHILDES' Manchester corpus (Theakston et al., 2000). Pronominal relatives (*e.g.*, '*the girl you like*') show a similar increase, and occur approximately as frequently. And prepositional phrases increase in frequency slightly more dramatically; they seem to occur approximately twice as often as relatives.[6]

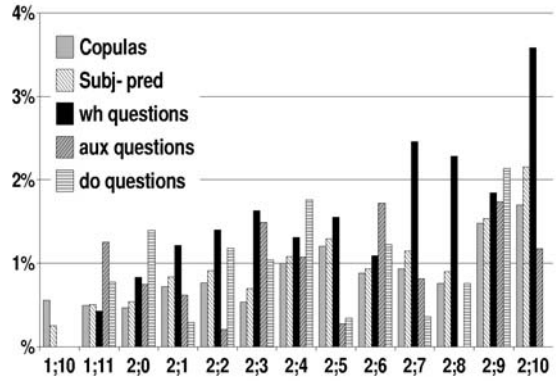


Figure 4: The percentage of NPs that contain relative clauses, for each month, averaged over all twelve children in the Manchester corpus.

The difference between the distribution of types in child-directed speech and speech between adults is also potentially significant. Child-directed speech contains a much greater proportion of questions — estimated at about one third of the child's input (Hart and Risley, 1995; Cameron-Faulkner et al., 2001) — and thus there is more of a balance between types. This may be critical in establishing the multiple roles that, *e.g.*auxiliaries, can take on; and also to reserve representational space for the the large variety of question forms. Figure 5 shows the percentages of copula constructions, subject-predicate forms (*e.g.*, transitives and intransitives), and *wh*-, *do*-, and *aux*-questions for representative months near the beginning, middle, and end of the time period covered by the Manchester corpus.

And finally, *aux*-questions in the child's input not only lack relative clauses in subject position, but are limited in a way that both predicts this absence, and potentially allows for the correct generalization to be formed. In child-directed speech, *aux*-questions with a determiner in the subject noun phrase — like '*Is the boy crazy?*' — are almost never used; the *aux*-questions in child-directed speech overwhelmingly use proper names, pronouns, deictics, *e.g.* '*Is that ...*', and other such forms which do not provide the correct context for a relative clause. Thus, given the low frequency of relative clauses in general, one should expect them to almost never occur in subject position.

These are ideal conditions for an SRN. The target generalization is supported by the appearance of relative clauses in all other positions in which noun phrases occur, and making the generalization incurs little cost since the context in which the generalization applies seldom occurs. If this were not the case, and questions like '*Is the boy crazy?*' were common, then the generalization would be threatened — each such occurrence would produce a false prediction which backpropogation would attempt to eliminate.

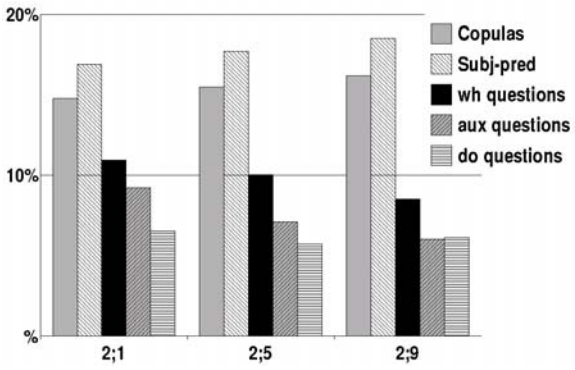


Figure 5: The percentage occurrence of various forms, at three stages, averaged over all children.

Motherese and the Generalization

Training sets generated on the basis of this analysis were used to determine if an SRN would generalize to predict (1), but not (2) from input of this sort. As before, the training sets contained *aux*-questions of the form 'AUX NP ADJ?'; but here the 'A_i NP B_i' forms were eliminated, and copula constructions, subject-predicate forms, and *wh*- and *do*-questions were added. The prohibition on NPs with relative clauses in *aux*-questions extended also to *wh*- and *do*-questions — *i.e.* NPs with relative clauses could occur in object position in these forms, but not in subject position. Thus these training sets also contained no evidence of the sort assumed to distinguish the structure-dependent hypothesis. Some examples from these training sets are given in Figure 6. The proportions of these general types, and the frequency of relative clauses and prepositional phrases, were manipulated in each portion of the training set to match with successive portions of the Manchester data — *e.g.*, the type distributions can be read directly from figure 5. And, as per the observation of the previous section, noun phrases in *aux*-questions were restricted to be, almost exclusively, names. The three training sets again consisted of 50,000

[6] A precise count of the prepositional phrases has not been made — in part because of the lesser significance to the current research issue, and in part because it is considerably more problematic to determine whether or not a prepositional phrase is within a noun phrase. But, (Cameron-Faulkner et al., 2001) analyzed a sample from this same corpus, and they report that prepositional phrases make up about 10% of all fragments, which may be indicative of their general frequency.

Mummy is beautiful.	*is Mummy beautiful?*
the little boy bites.	*is the little boy nice?*
the dog likes Mummy.	*is the dog hungry?*
does Mary smoke?	.
who likes Mary?	.
who does Mary like?	.
who likes the cat on the mat?	
who does the girl at the shop like?	
does the cat on the mat scratch?	
does the little girl like the boy who is smiling?	

Figure 6: Examples of the various types of utterances generated by the artificial grammar.

examples each; and again the network was trained for 10 epochs on each set, and was tested with the structures in (1) and (2) after each epoch.

Figures 7 and 8 chart the sum-squared error for (1) and (2) after each stage of training. As the figures show, the network succeeds in generalizing to predict (1), and generates significant error — and progressively larger error — at several points, when presented with (2).[7] The reasonably small error generated by the network when presented with '*who*' in the context of '*is the boy* _' shows that the relativizer is predicted. And the contrast in the errors generated by the subsequent presentation of either '*is*' or '*smoking*' shows clearly that the network has learned to predict an AUX after a relativizer, rather than entertaining the possibility of it's extraction, as in (2). Note, as well, that this contrast is monotonically increasing — at no point in training does the network predict a participle to follow the relativizer. And, for (1), the network's error is quite low for each successive word, including the presentation of the adjective after the participle, despite that '... PARTICIPLE ADJ ...' never occurs in the training sets. In contrast, for (2), as well as the error produced by the presentation of '*smoking*', the network also generates a substantial error upon the subsequent presentation of '*is*'; And though when presented with '*is the boy who smoking is* ' the network successfully predicts an adjective, the success is illusory: when subsequently presented with '*crazy*' the network's predictions are somewhat random, but a period is predicted more strongly than a question mark.

The network does, however, have some difficulties with this input. Although the grammar restricts relative clauses to the form 'REL AUX VERB*ing*', the network persists in predicting noun phrases and adjectives after the auxiliary — presumably because the '*is*' that occurs in initial position in *aux*-questions, followed by a noun phrase, and the '*is*' in declaratives, followed by an adjective, are relatively more frequent in the data than the '*is*'

[7]The SRN responsible for these results incorporates a variant of the developmental mechanism from (Elman, 1993). That version reset the context layer at increasing intervals; the version used here is similar, but does not reset the context units unless the network's prediction error is greater than a set threshold value.

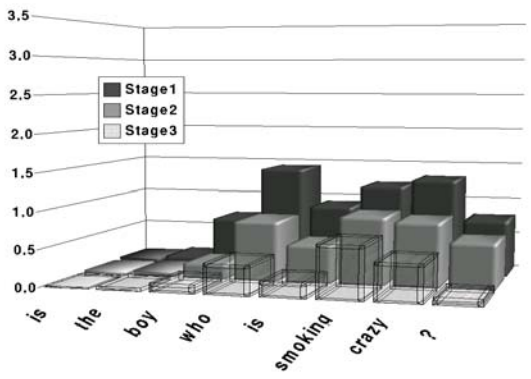


Figure 7: The sum-squared error after each word of the test sentence "*Is the boy who is smoking crazy?*" at the end of each stage of training.

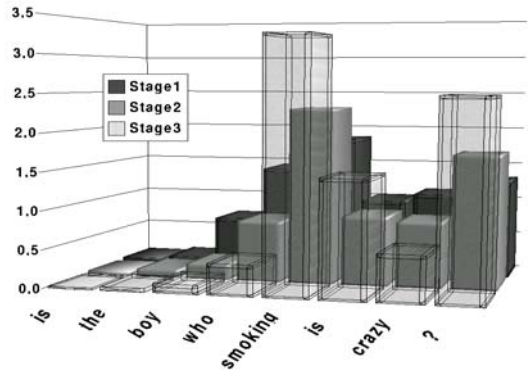


Figure 8: The sum-squared error after each word of the test sentence "*Is the boy who smoking is crazy?*" at the end of each stage of training.

in relative clauses. These erroneous predictions, however, gradually erode. And it is worth noting that they would be correct for a more realistic grammar.

The error associated with the adjective following the participle most likely has a similar source. Relative clauses occur only in either sentence final position, or preceding an auxiliary or a verb; thus the network initially expects participles to be followed by either a verb, a period, a question mark, or most prominently, an auxiliary. Again the problem is somewhat persistent, but is gradually resolved; by the end of the third stage such predictions, though remaining, are substantially weaker than the correct predictions — thus, arguably, not truly problematic. And it is plausible that such errors would not arise were the grammar to be made yet more realistic. The grammar used here contained little variation in terms of either NP types, syntactic structures, or lexical items, and thus generalizations were based on a quite limited set of distributional cues. Lifting the artificial limitations on the grammar might also help to eliminate such errors: questions like

'*what's the lady who was at the house called?*' — in Manchester's *ruth28a.cha* — are not only evidence of the sort assumed not to be available, but also data which discourage these sorts of false predictions.

But, such errors are also potentially meaningful. The most prominent and persistent of the errors is the prediction of an auxiliary following the participle, *i.e.*, '*is the boy who is smoking* <u>*is*</u> ...'; in fact an auxiliary is predicted as a possible continuation after any NP, *e.g.*,'*is the boy* <u>*is*</u> ...'. And this is an error that children make as well (Crain and Thornton, 1998).

Discussion

The objective here was to provide a proof that the structure of aux-questions is learnable from the input available to children. To make the results convincing, we have been careful to avoid providing the network with input that could be controversial with respect to its availability, and have represented the input in a way that encodes no grammatical information beyond what can be determined from its statistical regularities.

The fact that a neural network generalizes to make the correct predictions from input represented in this way, and modeled on child-directed speech — but limited to contain no data of what has been considered the relevant sort — shows that poverty of the stimulus arguments must give greater consideration to the indirect evidence available to the child. The statistical structure of language provides for far more sophisticated inferences than those which can be made within a theory that considers only whether or not a particular form appears in the input. And there is a growing body of evidence that children, not only neural networks, make use of the statistical properties of the input in acquiring the structure of language (*e.g.* Aslin et al., 1998; Gomez and Gerken, 1999). Thus learnability arguments cannot ignore those properties.

But discovering what those properties are, and determining their potential worth in language acquisition is difficult. This work shows that neural networks provide a means of dealing with this problem. As demonstrated here, neural networks can be used to assess just how impoverished the stimulus really is, and so can be invaluable to linguists in establishing whether or not a property of language can reasonably be assumed not to have been learned.

References

Aslin, R., Saffran, J., and Newport, E. (1998). Computation of conditional probability statistics by 8-month old infants. *Psychological Science*, 9:321–324.

Cameron-Faulkner, T., Lieven, E., and Tomasello, M. (2001). A construction based analysis of child directed speech. forthcoming.

Chomsky, N. (1968). *Language and Mind*. Brace & World, New York.

Chomsky, N. (1975). *Reflections on Language*. Pantheon Books, New York.

Christiansen, M. and Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, 23(2):157–205.

Cowie, F. (1998). *What's Within? Nativism Reconsidered*. Oxford University Press.

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences*, 14:597–650.

Crain, S. and Nakayama, M. (1987). Structure dependence in grammar formation. *Language*, 63:522–543.

Crain, S. and Thornton, R. (1998). *Investigations in Universal Grammar: A Guide to Experiment's on the acquisition of Syntax and Semantics*. MIT Press.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, 14:179–211.

Elman, J. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7:195–225.

Elman, J. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, 48:71–99.

Elman, J. (1998). Generalization, simple recurrent networks, and the emergence of structure. In Gernsbacher, M. and Derry, S., editors, *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, Mahway, NJ. Lawrence Erlbaum Associates.

Gomez, R. and Gerken, L. (1999). Artificial grammar learning by one-year-olds leads to specific and abstract knowledge. *Cognition*, 70:109–135.

Hart, B. and Risley, T. (1995). *Meaningful Differences in the Everyday Experiences of Young Children*. Paul H. Brookes, Baltimore, MD.

Morris, W., Cottrell, G., and Elman, J. (2000). A connectionist simulation of the empirical acquisition of grammatical relations. In Wermter, S. and Sun, R., editors, *Hybrid Neural Systems*. Springer Verlag, Heidelberg.

Piatelli-Palmarini, M. (1980). *Language and Learning: The debate between Jean Piaget and Noam Chomsky*. Harvard University Press, Cambridge, MA.

Pullum, G. and Scholz, B. (2001). Empirical assessment of stimulus poverty arguments. *The Linguistic Review*. to appear.

Rohde, D. (1999). Lens: The light, efficient network simulator. Technical Report CMU-CS-99-164, Carnegie Mellon University, Department of Computer Science, Pittsburgh, PA.

Sampson, G. (1989). Language acquisition: Growth or learning? *Philosophical Papers*, 18:203–240.

Slobin, D. (1991). Can Crain constrain the constraints? *Behavioral and Brain Sciences*, 14:633–634.

Theakston, A., Lieven, E., Pine, J., and Rowland, C. (2000). The role of performance limitations in the acquisition of 'mixed' verb-argument structure at stage 1. In Perkins, M. and Howard, S., editors, *New Directions in Language Development and Disorders*. Plenum.

Ties That Bind: Reconciling Discrepancies Between Categorization and Naming

Kenneth R. Livingston (livingst@vassar.edu**)**
Department of Psychology and Program in Cognitive Science
Vassar College, 124 Raymond Avenue, Box 479
Poughkeepsie, New York 12604

Janet K. Andrews (andrewsj@vassar.edu**)**
Department of Psychology and Program in Cognitive Science
Vassar College, 124 Raymond Avenue, Box 146
Poughkeepsie, New York 12604

Patrick Dwyer (padwyer@vassar.edu**)**
Department of Psychology
Vassar College, 124 Raymond Avenue, Box 3532
Poughkeepsie, New York 12604

Abstract

We present the results of a study designed to show that dissociations between lexical and similarity-based boundary partitions for a set of items can be produced in the laboratory. This is achieved by an incremental process of learning to assign a category label to items increasingly far removed (in similarity space) from the center of that category and approaching a different category. This process occurs in parallel with a compression effect in psychological similarity space such that increasingly distant items labeled as members of category A nonetheless come to be viewed as more similar to category B (the category to which they are in fact closer in pre-category learning similarity space) than they are by people who have not learned the category distinction.

Introduction

Although patterns of similarity are related in a more complex way to categorization and concept learning than was once thought, the evidence is mounting that such relationships are crucial to the partitioning of a set of items into categories (Medin, 1989; Medin, Goldstone, and Gentner, 1993; Goldstone, 1994a; 1994b). Among the most interesting recent findings is the discovery that the psychological similarity space usually assumed in the effort to measure relationships among category members is not static, but may actually undergo a change in its metric properties during the process of category learning. Thus, for example, Goldstone and his colleagues have found repeatedly that people who have learned to categorize a set of items make more reliable discriminations between pairs of items that cross the category boundary than people who have not learned to categorize them (Goldstone, 1994a; 1994b) This expansion or acquired distinctiveness effect seems to involve a stretching of the psychological similarity space in the region of the category boundary, with the result that smaller changes along the category relevant dimensions are sufficient to produce a just noticeable difference (JND).

Category clusters might also be formed by a process of compression in the region of similarity space that contains a set of items, with the result that a larger change is required to produce a JND. A number of researchers have found evidence for this process as well (Kurtz, 1996; Livingston, Andrews, and Harnad, 1998). Either of these processes, expansion or compression of the similarity space, is sufficient to produce a categorical distinction between sets of items, and thus looks promising as a general mechanism for concept learning. On such an account, a concept is formed when regions of psychological similarity space are warped so as to create a relatively more compact representational structure that can be more readily manipulated as a unit.

This description of a general mechanism for category learning is called into question, however, by recent findings concerning the relationship of language labels to category similarity structures. Malt, Sloman, Gennari, Shi, and Wang (1999) asked native speakers of English, Spanish, and Chinese to name each member of a set of sixty containers. They then asked these same people to sort these sixty items into groups based on their overall similarity. Perhaps not surprisingly, language communities differ in their lexical partitioning of the set of items. Chinese speakers partition the set using five words, English speakers use 7, and Spanish speakers use 15. The variation in grain is not the result of simply forming a greater number of subcategories in,

say, Spanish, of the same larger category boundaries found in Chinese; there is substantial non-shared variance in these lexical groupings.

The real surprise comes when one examines the clustering of items that occurs during the sorting task. Malt, et al., found that the sorts were very similar across language communities, in spite of the disparity in lexical boundaries for the set. There seems to be a dissociation of lexical boundaries from category boundaries based on similarity, and this creates something of a theoretical problem for the account of category learning given above. Indeed, it constitutes a puzzle for any theory of conceptual structure that suggests that there is a central tendency in the representation of the members of a category. The widely held belief that terms derive their meanings from their links to coherent concepts seems inconsistent with this result.

Unfortunately, we have no information about the kinds of judgments that people might have made of Malt, et al.'s set of containers *prior* to learning to name and categorize them. We therefore have no way of knowing how this odd state of affairs came to pass, nor whether it really constitutes a disconfirmation of the claim that psychological similarity space warps in a coherent fashion during category learning.

How might an item come to have a name other than that of its nearest neighbors with whom it shares a region in similarity space? One possibility discussed by Malt, et al. (1999) is that a series of intermediate cases could be introduced, one at a time, each inheriting the name of its nearest neighbor, and each more remote from the category to which the name was initially attached. If the chain of items spans the boundary between two categories, one could wind up with instances that are labeled as members of category A, the point of origin for the chain, even while having more in common perceptually with the members of category B. The dissociation of lexical and similarity groupings occurs because the former is based on nearest exemplar pairings introduced incrementally, while the latter is based on the warping of similarity space described above.

In order to test this hypothesis, we constructed a set of stimuli whose distribution in similarity space allowed for partitioning into two categories while leaving a set of items in the space between these groupings to allow the building of a naming chain from one to the other. Success at building such a chain would constitute an existence proof for this process, and comparison of data from people who learned to categorize the set with data from people without category or name-learning experience would allow us to determine whether this process alters the character of any warping of the similarity space.

Method

Participants

Participants were seventy-eight Vassar College undergraduates who were given course credit in an introductory psychology course. Twenty people participated in preliminary research to select an appropriate stimulus set. The remaining fifty-eight people participated in the experiment reported here, twenty-two of them in a control group, and eighteen in each of two experimental groups.

Stimuli

Ten members of the Nemipteridae family of fish (threadfin breams) and ten members of the Labridae family (wrasses) were selected for preliminary analysis from Burgess, Axelrod, and Hunziker (1997). Each stimulus was color photocopied, glued onto a blank card (5.1 by 10.8 cm), laminated, and then randomly assigned a number from 1 to 20, which was printed on the back of the card. Following examination of the two-dimensional solution to a multidimensional scaling analysis (MDS) of the similarity judgments of twenty people on all 190 possible pairs (see Procedure below), the set was culled to remove outlying cases. These were then replaced with items that occupied the central region of the 2D space, which appeared to be defined by the dimensions of degree of body striping, and the ratio of body width to length. These replacement items were selected without regard to membership in the two original categories.

Procedure

Participants were assigned either to a control condition (N=22) or to one of two learning conditions (N=18 each group). Participants in the control condition were asked to judge the degree of similarity between all possible pairs of the twenty stimuli (190 pairs). The participant was seated at a table upon which the stimuli had been placed face up, and was allowed to inspect the entire stimulus set for one minute. The stimuli were then turned over, revealing the numbers on the back, and the experimenter began naming pairs in a previously determined random order. The participant was instructed to turn over each corresponding pair, look at the two items, and verbally rate the similarity of the stimuli on a 9-point scale from 1 (most similar) to 9 (most different). Ratings were provided to one decimal place.

For control group participants, the similarity judgment task was the only task required. All control group participants were run through the procedure in a block before work began with the experimental group. This allowed us to complete an MDS analysis of the data from this group to confirm the pattern of similarity relationships in the set and the choice of stimuli for building the chain between categories. As expected, the MDS analysis of the control group data revealed that

stimuli clustered in two core groups consisting of eight and seven items, respectively. The five remaining stimuli were intermediate cases, four of which formed a chain between the two larger groups. One intermediate stimulus that was not part of the chain was treated as a neutral stimulus, and did not appear in the training task. It served as the means to a test for demand effects in the data set (see Results, below).

This analysis of the pre-categorization similarity space allowed us to design the stimulus sets for the learning group. Learning participants were first taught to categorize the core set of fifteen stimuli. Stimuli were presented individually, in blocks of fifteen that included all members of both categories. The participant was asked to label each picture as either *gracilia* or *aurora*. The experimenter gave immediate feedback, recorded the response, and then presented the next picture. Order of presentation within each trial block was random. Training continued until the participant met the criterion of two consecutive errorless trial blocks (30 stimuli), or a total of 20 trial blocks had passed, whichever came first.

Once category training on the core stimuli was complete, the first stimulus in the chain was introduced into the subsequent trial block. Which stimulus this was depended on the direction in which the chain was being built. For half of the study participants, the chain was built from *gracilia* to *aurora* (the G-root group) and for half it was built in the opposite direction (the A-root group). Assignment to these subgroups was random. The first chaining stimulus introduced was the one closest to the root category in the MDS space derived from the control group data.

Once this new stimulus was introduced into the set, training continued as before, except that trial blocks now contained one additional stimulus. Order of presentation was still random, except that the new chaining stimulus was always presented immediately after its nearest neighbor in the root category. Training using this newly expanded set continued until the participant met the criterion of two consecutive trial blocks in which all chaining stimuli presented up to that point had been correctly categorized, or a total of 5 trial blocks (after introduction of the new item), whichever came first. Once the participant met criterion for one chaining stimulus, the next stimulus in the chain was introduced into the subsequent trial block. The new chaining stimulus always followed the previously learned one, which appeared randomly in the trial block along with the 15 core stimuli and any other chaining stimuli. Training was halted when the participant met the training criterion after the introduction of the fourth and final item in the chain.

Once training was completed, the participant took a short break before completing the similarity judgment task. Procedures for this task were exactly as for the people in the control group, and are described above.

Results

Six people in the learning group did not meet the criterion for successful learning by the conclusion of training. All six of these people were in the G-root group. Data for those who failed to reach criterion were excluded from all analyses. Mean similarity ratings were calculated separately for the control and learning groups, and, within group, mean similarities were calculated for *aurora-aurora* (A-A) pairs, *for aurora-gracilia* (A-G) pairs, and for *gracilia-gracilia* (G-G) pairs. Separate analyses were then performed for the two different chaining groups, using the same control group data for comparison in both cases.

In order to determine whether category learning produced compression and/or expansion effects in the A-root group, we performed a 2 (group: control vs. learning) by 3 (pair type: A-A, A-G, G-G) analysis of variance (ANOVA) on mean similarity ratings with repeated measures on the second variable. The analysis revealed a significant main effect of pair type, $F(2,76) = 149.506$, $MSE = 63.152$, $p < .0001$; and a significant interaction effect, $F(2,76) = 9.952$, $MSE = 4.202$, $p < .001$. (See Figure 1). For the G-root group, the same 2 (group: control vs. learning) by 3 (pair type: A-A, A-G, G-G) ANOVA with repeated measures on the second variable revealed significant main effects of group, $F(1,32) = 9.488$, $MSE = 22.437$, $p < .005$, and pair type, $F(2,64) = 57.551$, $MSE = 20.376$, $p < .0001$. The interaction effect was not significant. (See Figure 2.)

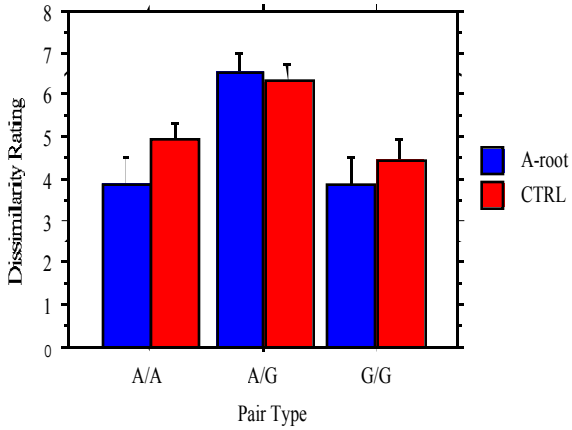


Figure 1. Comparison of the mean similarity ratings of the control group and the group who learned a chain of items rooted in the *aurora* category. Interaction of group with pair-type is shown.

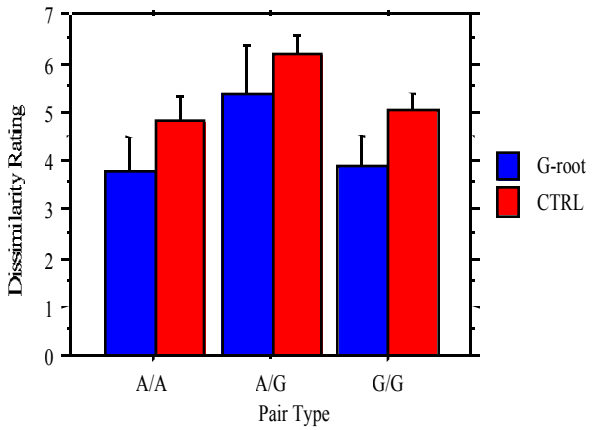


Figure 2. Comparison of the mean similarity ratings of the control group and the group who learned a chain of items rooted in the *gracilia* category. Interaction of group with pair-type is shown.

Thus, compression occurred in both learning groups, i.e., item pairs were judged to be more similar relative to control group ratings, particularly for the within-category (A-A and G-G) pairs. No expansion occurred, i.e., between-category pairs (A-G) were not judged to be less similar by the learning groups than by the control group.

Neutral Stimulus Analysis

Goldstone, Lippa, and Shiffrin (2001) have devised an ingenious technique for detecting the presence of demand effects in data of this kind. The analysis works by comparing the pattern of changes in similarity relationships among items in a learning set relative to a neutral item not included in training. If the similarity judgments of pairs including the neutral item change in ways that are predictable from the compression or expansion effects that occur for categorized items, this must be due to actual changes in the underlying similarity space and not demand effects, since the neutral item was never categorized.

For our data set, the absolute difference for each participant between all possible within-group and between-group pairs of pairs involving the neutral stimulus was calculated, averaging separately across each group. Separate 2 (group: learning vs. control) by 2 (pair of pair type: within vs. between) ANOVAs with repeated measures on the second variable were conducted for both the A-root and the G-root groups. The pattern of results was the same as that reported above. There was thus no evidence that the observed compression was due to a demand effect for either group.

MDS Analysis

In order to better understand the nature of the changes taking place in psychological similarity space we performed a multi-dimensional scaling analysis of the mean similarity ratings of all three groups (control, A-root, and G-root). The full matrix of mean similarities from each of the three groups was entered into an INDSCAL analysis. The two-dimensional solution provides a relatively good fit to the data (R-squared = .872), and is plotted in Figure 3. The locations of all twenty stimuli for each of the three groups (control, A-root, and G-root) are depicted, and the pattern of changes in similarity is shown by arrows connecting three points. The point at the tail end of the arrow represents the location in the space of the stimuli as judged by the control group. The middle point shows the locations of the twenty stimuli as judged by the G-root group, while the points at the tips of the arrow heads show the locations of the twenty stimuli as judged by the A-root group. The graph gives a sense of the compression that occurs in the similarity space as categories are learned. Note that the greater relative proximity of items following compression of the similarity space has the effect of making them a more easily identified grouping, even though there is no statistically significant increase in mean inter-item differences across the category boundary. Most importantly, the graph shows how the chained items move toward the central tendency of their nearest neighbor cluster, even when the label training ties them to the more remote cluster.

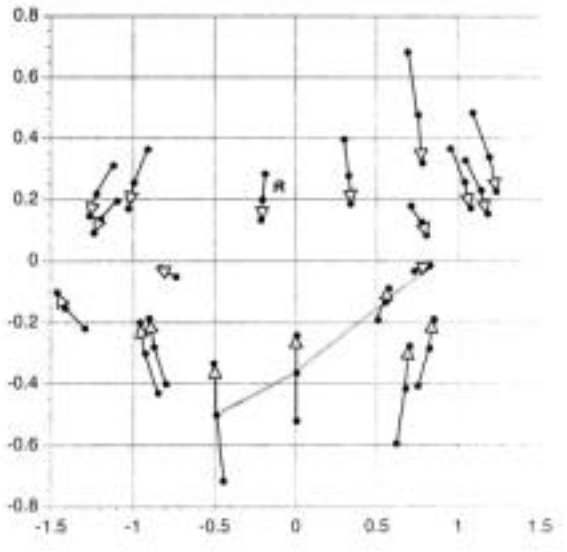


Figure 3. Shows the two-dimensional MDS analysis (INDSCAL) of similarity relationships among the twenty stimulus items for all three groups. The arrow labeled R is for the neutral item (see text). The four arrows linked by a shaded line are the chaining items.

The remaining points on the right side of the graph are the core items in the *aurora* category, while those on the left are the core items in the *gracilia* category. The space of similarities for the control group is represented by closed circles found at the tails of the arrows. The circles at the tips of the arrow heads represent the space of similarities for the experimental group that learned a chain that begins in the *aurora* category and extends toward the *gracilia* category. The filled circles found roughly in the middle regions of the arrows represent the space of similarities for the group that learned a chain that begins in the *gracilia* category and extends toward the *aurora* category.

Discussion

The domain of words must map in some predictable way to the domain of concepts if there is to be anything to the story that says words have meaning in virtue of the concepts to which they are linked. This is why Malt, et al.'s (1999) finding of an apparent dissociation between lexical boundaries and category boundaries in similarity space is so provocative. The experiment reported here demonstrates one process by which names might become attached to items that are remote in similarity space from the core of the concept to which the name typically refers. Note that although we did not measure typicality in this study, one further prediction would be that chained items should be seen as atypical of the category they name.

More important than the demonstration of a procedure for producing this dissociation are the data showing how this effect is related to the warping of similarity space previously shown to occur during category learning. This effect was clear in comparisons of the A-root group (people who learned a chain that begins with an item well within the region of the *aurora* group) with the control group. Analysis of variance revealed the same pattern of within-category compression found in previous research (e.g., see Kurtz, 1996; Livingston, et al., 1998). The effect is less clear in the G-root group. Compression occurs in this case, but it occurs for between category pairs as well as for within-category pairs. Obviously, the direction in which we tried to build chains made a difference, an observation further confirmed by the fact that all of the study participants who failed to reach our learning criterion were in the G-root chaining group. The nature of the difference between our two experimental groups, and its consequences for similarity judgments, can be seen in Figure 3.

First, notice that the chains differ in how deeply rooted they are in their originating categories. The *aurora*-based chain begins with an item well inside the region of similarity space that encompasses the category, and it does not extend very deeply into the region occupied by the *gracilia* category. Exactly the opposite is true for the G-root chaining group. Thus, even when people are learning to label the last two items in the chain leading deeply into *aurora* territory as *gracilia*, the region of space that they occupy is being compressed still further around the central tendency of the *aurora* category. It is therefore not surprising that one sees evidence of what counts as between-category compression for this group, because those last two items in the chain are considered *gracilia* for purposes of this analysis.

Looked at from this lexical perspective, the result seems straightforward enough, but the effect is far more interesting for what it tells us about how category learning warps similarity space. Notice that the lexical chaining effect seems to have relatively little effect on the direction in which similarity space is compressed during learning. The overall magnitude of the effect is different in the two cases, an effect that is likely the result of the fact that the task is more difficult and so produces slower and less robust learning in the G-root group, but the space warps in the same way in both cases. This warping, based on the pattern of structure and variation in the whole set of items, overrides local tendencies associated with rogue items chained in from elsewhere. As can be seen in Figure 3, the result is that the chained items move in the direction of the nearest region of compression, even while they are being labeled as members of the more remote cluster. The similarity-based warping of the representational space occurs independently of labeling. Thus are lexical and similarity-based category dissociations produced.

If this account is correct, several further predictions follow. We have already mentioned the predictions for typicality. Over long periods of time, we would also expect chains that stretch far from their roots would become unstable and break, with items at the ends of those chains receiving new labels more in keeping with those of their similarity-space neighbors. Linguistic analysis of patterns of lexical evolution should reveal such phenomena in the history of any language. We would also expect this pattern to be most common for artifact categories, where genuinely new instances are introduced with some frequency, and the perceptual and functional feature landscape is quite fluid. We are currently conducting a replication and extension of the study reported here using artifact categories and additional control groups. Finally, these results have deeper implications for the nature of the relationship between systems for representing lexicons, at least and systems for representing category information. These two systems must remain in register to some extent, but it is clear that each is sufficiently modular with respect to the other to permit some rather remarkable dissociations to develop in very short order.

Acknowledgments

Our thanks to the Undergraduate Research Summer Institute at Vassar College, and to Maria Jalbrzikowski and Paul Francaviglia for their assistance in the conduct of this experiment.

References

Burgess, W.E., Axelrod, H.R., & Hunziker, R. (1997). Dr. Burgess's *Mini-Atlas of Marine Aquarium Fishes Mini-Edition*. Neptune City, NJ: TFH Publications, Inc.

Goldstone, R. L. (1994-a). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General, 123,* 178-200.

Goldstone, R. L. (1994-b). The role of similarity in categorization: Providing a groundwork. *Cognition, 52*, 125-157.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. (2001). Altering object representations through category learning. *Cognition, 78,* 27-43.

Kurtz, K. J. (1996). Category-based similarity. In G. W. Cottrell (Ed.) *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*, 290.

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 732-753.

Malt, B.C., Sloman, S.A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, *40*, 230-262.

Medin, D.L. (1989). Concepts and conceptual structure. *American Psychologist, 44*, 1469-1482.

Medin, D.L., Goldstone, R.L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100,* 254-279.

Effects of multiple sources of information on induction in young children

Yafen Lo (yafen@rice.edu)
Department of Psychology
Rice University
270 Sewall Hall
Houston, TX 77005, USA

Vladimir M. Sloutsky (sloutsky.1@osu.edu)
Center for Cognitive Science & School of Teaching & Learning
Ohio State University
21 Page Hall, 1810 College Road
Columbus, OH 43210, USA

Abstract

This report considers differences in induction of biological properties between children and preadolescents based on differences in stimuli processing in these two groups. Two studies test predictions that young children, but not preadolescents, base their inductive inference on aggregating information from different sources rather than relying on a single source of information. In both experiments 4-5 year-olds, 7-8 year-olds, and 10-11 year-olds were presented with an inductive task. In Experiment 1, linguistic labels were fully crossed with relationship information, whereas in Experiment 2 perceptual similarity information was fully crossed with relationship and labeling information. While 10-11 year-olds relied exclusively on inheritance across experiments, 4-5 year-olds relied on an aggregate of multiple sources of information, and 7-8 year-olds fell between these two extremes. In addition, while the relative weight of inheritance on inferences increased with age, the weights of other information sources decreased. These results support the hypotheses suggesting that between 8 and 10 years of age children undergo a developmental shift from a holistic feature-integration induction to knowledge-based induction based on a single most predictive source.

Introduction

Inductive inference, or extending knowledge from known to novel instances, is ubiquitous in human cognition. For example, if one learned that a particular cat uses acid-based enzymes for digestion, one would expect other cats also to use acid-based enzymes for digestion, without having factual knowledge of digestion in cats.

The simplest case of inductive inference is induction over individuals, when attributes or relations are generalized from a single entity to another single entity, with both entities beings members of the same category (e.g., *This Bird has biological property X*, *therefore that Bird has biological property X*). There is a large body of research demonstrating the ability of young children to perform specific induction of biological properties (e.g., Gelman, 1988; Gelman & Markman, 1986, 1987; Gutheil, Vera, & Keil, 1998; Johnson & Solomon, 1997; Rosengren, Gelman, Kalish, & McCormick, 1991; Sloutsky & Lo, 2000; Sloutsky, Lo, & Fisher, in press; Solomon, Johnson, Zaitchik, & Carey, 1996; Springer, 1996; Springer & Keil, 1989). In addition, several lines of research have emerged in an attempt to determine what aspects, or what information cues, of compared entities children rely on when performing such induction.

One important aspect of previous research is that the majority of tasks pitted one information cue (or source of information) against another (e.g., appearance versus label, or appearance versus inheritance). This was necessary to establish the predictive value of each cue relative to a competing cue. At the same time, people typically face stimuli comprising multiple sources of information bundled together, with several sources supporting induction, but each having different predictive value. In particular, while the Target may comprise a bundle of cues $C_{i-1}C_{j-1}C_{k-1}$ (a particular appearance, inheritance, and category label), one entity may comprise another bundle $C_{i-1}C_{j-1}C_{k-2}$ (e.g., sharing appearance and category label with the Target), whereas another entity may comprise $C_{i-2}C_{j-2}C_{k-1}$ (e.g., sharing only inheritance with the Target). For example, a baby boy shares inheritance with his mother, whereas he shares appearance (at least in terms of his size and outfit), gender, and linguistic label "baby boy" with his neighbor baby boy. Would people induce from one baby boy to another or would they induce from a baby boy to his mother? It is reasonable to expect that when performing induction, adults would rely on a single most predictive cue: age to predict sleeping patterns, sex to predict gender development, and inheritance to predict blood type. However, it remains unclear how children perform induction across entities sharing multiple sources of information. Do they perform induction by relying on a single cue or do they aggregate information from different cues? In addition, if they rely on a single source of information, does the

importance of this source change in the course of development? Or if they rely on multiple sources of information, does the relative importance of each source change in the course of development?

Answers to these questions depend on how young children process multiple information cues. If they process each cue separately, one cue at a time, then due to working memory limitations (see Hitch & Towse, 1995, for a review), they should invariably rely on one most salient cue. On the other hand, if they process complex information in a holistic manner without attending to specific dimensions of stimuli (Shepp, 1978; Smith, 1989a, 1989b), they should rely on an aggregate of multiple sources. These different processing mechanisms may result in different developmental scenarios. If young children process cues separately, and development is a function of increasing working memory, then both young children and adults should rely on a single cue, with adults exhibiting larger flexibility in cue selection. For example, in the baby boy example, adults, but not young children, should use different cues when inducing gender development versus inducing the blood type. On the other hand, if, unlike adults, young children process cues holistically, then young children and adults should exhibit more profound differences, with young children relying on multiple sources of information, while adults relying on a single, most predictive source. Therefore, answers to the posed questions are important for understanding of developmental mechanisms of knowledge generalization and inductive inference, as well as general principles of the development of stimuli processing.

The overall experimental approach is as follows. The task consisted of presenting participants with triads of pictures. Each triad included a Target (a Baby animal), Test Stimulus A (a neighbor animal "who played with the baby") and Test Stimulus B (an animal "who gave birth to the baby"). Within each triad, participants were asked to generalize an unobservable biological property from the Test stimuli to the Target (e.g., blood color). In Experiment 1 the Target and the Test stimuli received labels and inheritance information, while perceptual similarity was kept constant. For half of the triads the Target shared a linguistic label with Test A while on the other half it shared a label with Test B. In Experiment 2, in addition to relationship information and labels, participants were also presented with perceptual information (with one Test stimulus being perceptually similar to the Target, while the other one being dissimilar), with the three attributes fully crossed.

Experiment 1

Method

Participants Participants were 45 children and preadolescents recruited from one daycare center and one elementary school located in middle class suburbs of Columbus, Ohio. There were three age groups, with 15 participants in each: (1) 4-5 year-olds (5 boys and 10 girls; M = 4.5 years; SD = .66 years); (2) 7-8 year olds (7 boys and 8 girls; M = 7.7 years; SD = .51 years); and (3) 10-11 year-olds (7 boys and 8 girls; M = 10.4 years; SD = .68 years). These participants were selected on the basis of returned parental consent forms.

Design and Materials The experiment had a mixed design with age as a between-subject factor and the information condition as a within-subject variable. The information condition had two levels: (1) Inheritance only information (when the Target shared only inheritance information and not labeling information with the Mother) and (2) Inheritance + Label information (when the Target shared both inheritance information and the label with the Mother). Note that in the Inheritance only condition the Target shared the label with the Friend, whereas in the Inheritance + Label condition, the Target had a label that was different from that of the Friend.

The order of Inheritance only and Inheritance + Label trials was counterbalanced across participants. Each participant was presented with eight stories (four stories in the Inheritance only condition and four stories in the Inheritance + Label condition).

Materials consisted of triads of line-drawing pictures with a fully shown Target animal, and Test A and Test B stimuli hidden behind trees, stories, biological properties, and auditorily presented linguistic labels. Each triad of stimuli had two labels, so that either Test A or Test B shared the label with the Target, whereas the other Test stimulus had a different label. To avoid confounds with existing knowledge about specific animals, we used only artificial labels, each consisting of a short two-syllable word (e.g., Jiga, Gapo, etc.). The labels were presented as count nouns (e.g., “look, this is a Jiga”). After each label was introduced, children were asked to repeat the label.

Procedure

The experiment was conducted in a single 10-15 minute session, during which participants were read four short stories, one story at a time. Each story constituted a trial that included three phases: stimuli presentation, comprehension/memory check, and inductive inference. Each participant was asked a total of 16 inductive inference questions with four questions for each of the four stories. Participants were tested individually in a quiet room by a female experimenter.

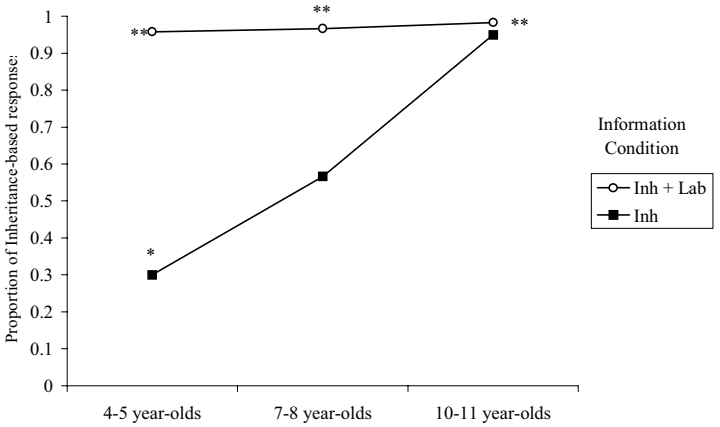
First, participants were read a cover story describing a baby animal who saw two adult animals playing in the forest. One of these animals was introduced as the one "who used to play with the baby," while the other was introduced as the one "who gave birth to the baby." The order of presentation of the Test stimuli was counterbalanced across the stories, and the order of introduction of attribute pairs was randomized across trials and across participants. Then participants were told that each of the Test stimuli (i.e., Mother vs. Friend) has a particular biological property (e.g., thick blood vs. thin blood) and asked which of these properties are likely to be shared by the Target (i.e., Baby).

Results and Discussion

Proportions of Inheritance-Based induction broken down by age group and information condition are presented in Figure 1. As shown in Figure 1, in the Inheritance only condition (Inh), the proportion of Inheritance-Based generalizations differed across the age groups. In the group of 4-5 year-olds this proportion was at 30% (below chance, Confidence Interval from 17.3% to 40%, $p < .01$) and in the group of 7-8 year-olds this proportion was at 57% (not statistically different from chance). At the same time, in the group of 10-11 year-olds the proportion of Inheritance-Based induction was at 95% (Confidence Interval from 86% to 99%, ps $< .001$). As opposed to Inheritance only condition, in the Inheritance + Label condition (Inh + Lab), the majority of participants of all age groups (over 95% in each group) responded in an Inheritance-Based manner.

Proportions of Inheritance-Based induction were subjected to a two-way (age group by information condition) repeated measures ANOVA. The analysis revealed significant main effects of age group, $F(2, 42) = 10.7$, $p < .0001$, and information conditions, $F(1, 42) = 42.3$, $p < .0001$, and a significant age group by information condition interaction, $F(2, 42) = 10.5$, $MSE = 4.5$, $p < .0001$. Post-hoc Bonferroni tests of the main effect of age indicated that children in the two youngest groups performed Inheritance-Based induction significantly less frequently than did children in the oldest group, all ps $< .05$. The second main effect (indicating that in the Inheritance + Label condition participants performed Inheritance-Based induction more frequently than in the Inheritance only condition) was largely driven by the interaction. To analyze the interaction, t-tests with Bonferroni adjustments for multiple comparisons were performed within each age group. The analysis pointed to significant differences between the Inheritance + Label and the Inheritance only conditions in the group of 4-5 year-olds and 7-8 year-olds, both ts > 5, ps $< .01$. At the same time, there were no such differences in the group of 10-11 year-olds, $t < 1$. Therefore, younger children were more likely to perform Inheritance-Based induction when both Inheritance and Label information supported such induction than when induction was supported by Inheritance information alone. At the same time, older children relied solely on inheritance information, while ignoring labeling information altogether.

Figure 1. Proportion of Inheritance-Based induction by age group and labeling condition



Note: ** Above chance, $p < .01$; * below chance, $p < .01$

However, Experiment 1, while presenting suggestive evidence, does not rule out an alternative explanation that 4-5 yea-olds perform induction across similarly labeled entities, while ignoring inheritance information altogether. To test this alternative, we conducted Experiment 2, where additional information cues were added to the design. If young children rely on multiple sources of information when performing induction, their induction should be a function of the number of information sources shared by compared entities. Another goal of Experiment 2 was to examine whether or not the relative importance of each source change in the course of development?

Experiment 2

Method

Participants Participants were 96 children recruited from two daycare centers, two elementary schools, and one middle school located in middle class suburbs of Columbus, Ohio. These participants represented three age groups each consisting of 32 children: (1) 4-5 year-olds (15 boys and 17 girls, $M = 4.8$ years; $SD = 0.63$ years); 7-8 year-olds (14 boys and 18 girls, $M = 7.6$ years; $SD = 0.74$ years); and 10-11 year-olds (15 boys and 17 girls, $M = 11.2$ years; $SD = 0.54$ years).

Design and Materials This experiment had a mixed design with age as a between-subject factor, perceptual similarity (i.e. Target perceptually similar to Test A vs. Target perceptually similar to Test B) as a between-subject variable, and information condition as a within-subject variable. The crossing of perceptual similarity and information conditions resulted in four cells. (1) Inheritance only (Inh) where the Target shared only inheritance information with the Mother. (2) Inheritance + Label (Inh + Lab) where the Target shared inheritance and the label with the Mother. (3) Inheritance + Perceptual Similarity (Inh + PS) where the Target shared inheritance and appearance with the Mother. And (4) Inheritance + Label + Perceptual Similarity (Inh + Lab + PS) where the Target shared inheritance, the label, and appearance with the Mother. All materials were identical to those used in Experiment 1, except that stimuli in the present experiment showed the Target, Test A, and Test B fully.

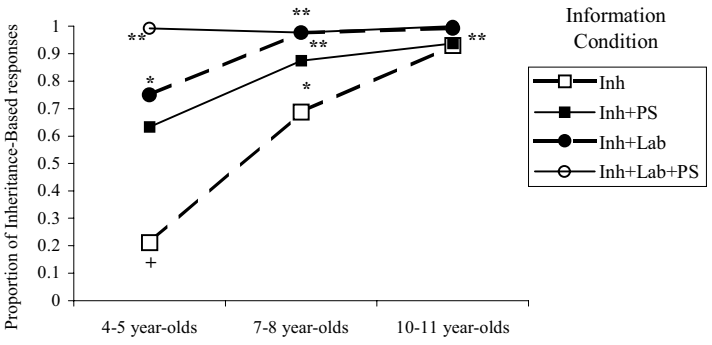
Similarities among stimuli were estimated in a calibrating experiment with 10 adult participants judging similarity between each pair of stimuli. Similarity scales ranged from 5 (very similar) to 0 (very dissimilar). The mean similarity rating for those pairs that we deemed dissimilar was .63 (*SD* = .43), while for those that we deemed similar it was 4.45 (*SD* = .23), $t(18) = 24.86$, $p < .0001$.

Procedure The procedure was identical to that in Experiment 1.

Results and Discussion

Proportions of Inheritance-Based responses, broken down by age group, similarity, and information conditions, are presented in Figure 2.

Figure 2. Proportion of Inheritance-Based induction by age group and information condition



Note: * Above chance $p < .05$; ** above chance $p < .005$; + below chance, $p < .005$.

As shown in Figure 2, participants of different age groups differed in their reliance on various information sources in the course of induction. In the group of 4-5 year-olds, the proportion of Inheritance-Based induction increased with the number of available information sources ranging from 20% in the Inh condition to 65-75% in the Inh + Lab and Inh + PS conditions, and to 99% in the Inh + Lab + PS condition. At the same time, in the group of 10-11 year-olds no such differences were observed: in this group proportions of Inheritance-Based induction were at ceiling across the information conditions. The group of 7-8 year-olds was between these extremes with differences among the conditions being larger than in the group of 10-11 year-olds, but smaller than in the group of 4-5 year-olds.

In short, data in Figure 2 suggest that preadolescents relied only on inheritance information, while ignoring other sources of information. At the same time, participants of the two younger groups relied on multiple sources of information. Another aspect of the findings, as indicated in Figure 2, is a sharp developmental increase in the importance of inheritance information.

Results presented in Figure 2 were subjected to a 3-way (age group * similarity condition * information condition) ANOVA with age group and similarity condition as between-subjects factors and labeling condition as a repeated measure. All main effects were significant. First, there was a significant main effect of age group, $F(2, 90) = 26.1$, $MSE = 4.3$, $p < .0001$, with 4-5 year-olds generalizing biological properties from the Mother significantly less frequently than children in the two older groups (63% vs. 88% vs. 96%), post-hoc Bonferroni tests, *p*s < .0001. Second, there was a significant effect of perceptual similarity condition, $F(1,90) = 15.1$, $MSE = 4.3$, $p < .0001$, with a tendency to perform an Inheritance-Based induction more frequently when the Mother looked similar to the Target (90% vs. 76%). Third, there was a main effect of Information condition, $F(1,90) = 57.4$, $MSE = 4.3$, $p < .0001$, with the overall tendency to perform Inheritance-Based induction more often in the Inheritance + Label than in the Inheritance only condition (95% vs. 71%).

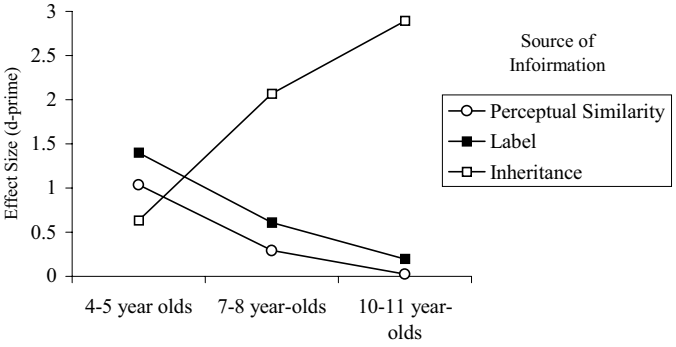
The two latter main effects, however, were largely driven by the two significant interactions. First, there was a significant age group by perceptual similarity interaction, $F(2, 90) = 6.8$, $MSE = 4.3$, $p < .003$, with significant effects of perceptual similarity on induction in the youngest group, both *t*s for both within-subject conditions > 3, *p*s < .01, but no such effects in the older groups (all ps > .27). Second, there was a significant age group by information condition interaction, $F(2, 90) = 3.3$, $MSE = 3.0$, $p < .0001$, with large (i.e., 54%) differences between the Inheritance + Label and the

Inheritance only conditions in the youngest group, t (31) = 6.8, $p < .0001$, and smaller (i.e., 29% and 6%, respectively) differences in the older groups, both ts > 2.3, ps < .05. No other interactions were significant.

To analyze interactions, t-tests, comparing means for each condition, were conducted within each age group. In the group of 4-5 year-olds, the proportion of Inheritance-Based generalizations exhibited the following differences (1) Inh (21%) < Inh + PS (64%) = Inh + Lab (75%) < Inh + PS + Lab (99%), all ts > 3, ps < .01, for differences. In the group of 7-8 year-olds, the proportion of Inheritance-Based generalizations exhibited the following differences: Inh (69%) < Inh + PS (88%) = Inh + Lab (98%) = Inh + PS + Lab (98%), all ts > 2.8, ps < .05, for differences. At the same time, in the group of 10-11 year-olds no significant differences among the conditions were found, all proportions ranging between 99% to 100%, all ps > .1.

Data presented in Figure 2 allowed us to estimate the relative contribution of inheritance information, shared label, and perceptual similarity to generalizing biological properties from the Mother. To do so, we calculated effect sizes for each of these sources of information, by dividing differences between marginal means for each of the sources (e.g., M perceptual similarity – M no perceptual similarity) by pooled standard deviations (Cohen, 1988). These estimated contributions of inheritance information, perceptual similarity and labeling broken down by age group are presented in Figure 3.

Figure 3. Relative contribution (effect sizes) of different sources of information to inductive inference by age group



Effect sizes exhibited the following patterns: while relative contributions of labels and perceptual similarity tend to decrease with age, the contribution of inheritance increased dramatically with age. In particular, effect sizes due to labels and perceptual similarity decreased from 1.4 and 1, respectively, in the youngest group to 0.2 and 0.02, respectively, in the oldest group. At the same time, effect sizes due to inheritance increased from 0.65 in the youngest group to 2.9 in the oldest group. In short, while for children of the youngest group all sources of information made sizable contributions to induction (all d-primes 0.65), participants of the oldest group (i.e., preadolescents) relied almost exclusively on inheritance information. At the same time, 7-8 year-olds were between these two extremes. In particular, for this group, inheritance information made a greater contribution than either perceptual similarity or labeling information, while the effect size due to labeling was still quite sizable (d-prime = 0.61).

In short, 4-5 year-olds exhibited maximal proportions of Inheritance-Based induction when all three sources of information supported this induction, 7-8 year-olds were at the maximum when at least two sources supported Inheritance-Based induction, and 10-11 year-olds were at the maximum even when only inheritance information was available.

General Discussion

The reported findings fit predictions well, supporting our contention that young children rely on multiple sources of information when performing induction. The larger the informational overlap between the Target and the Test stimuli, the more likely that a biological property would be generalized from the Test to the Target. For example, Inheritance alone contributed less than Inheritance + Perceptual Similarity or Inheritance + Label, which, in turn, contributed less than Inheritance + Label + Perceptual Similarity.

Findings of the reported experiments point to two important developmental changes: (1) increasing reliance on a single source of information and (2) increasing salience of inheritance information accompanied by decreasing salience of labeling and perceptual similarity.

The first change supports the developmental scenario in which processing develops from holistic to specific. As predicted by this scenario, young children tended to perform induction relying on multiple sources of information, whereas preadolescents, regardless of the number of sources of information, performed induction relying on a single source. Of course, it could be argued that the observed pattern of responses could stem solely from the second change – the increasing importance of inheritance information. However, previous research suggested that preadolescents do not focus on inheritance per se when performing induction, but they rather focus on a single attribute that they deem most predictive (Sloutsky & Lo, 2000; Sloutsky, Lo, & Fisher, in press), they consistently relied on the label information, considering it more reliable predictor than appearance information. Hence, it seems that the tendency to rely on a single source increases with age,

independently of the increasing salience of inheritance information.

The second change points to a decrease in the importance of less predictive sources of information (i.e., appearance and labels) and an increase in the importance of a more predictive source (i.e., inheritance). Both developmental changes suggest that between 8 and 10 years of age children undergo a developmental shift from a feature-integration induction to a single-feature, knowledge-based induction.

Current experiments also raise questions about the nature of young children's induction. If children's induction is driven by their intuitive theories, they should be able to attend separately to each predictor, and then to integrate information from different predictors. On the other hand, if the reliance on multiple features stem from their inability to selectively attend to each of the source and the inability to separate sources, this would be indicative that induction is not based on intuitive theories. This is because intuitive theories are beliefs about the world, and, therefore, they could not be products of low-level pre-attentive mechanisms. Current results cannot conclusively distinguish between these possibilities. This issue, however, could be addressed in future research directly examining separability of inheritance, labeling, and perceptual information in children and preadolescents.

In sum, this research suggests that when compared stimuli comprise multiple sources of information, 4-5 year-olds tended to rely on several sources when performing induction, 7-8 year olds relied mostly, but not exclusively on inheritance information, whereas 10-11 year-olds relied solely on inheritance information. Therefore, in the course of development, children undergo a transition from performing induction relying on multiple sources of information to performing induction relying on a single, most predictive source.

Acknowledgments

This research has been supported by grants from the James S. McDonnell Foundation and the National Science Foundation to the second author.

References

Cohen, J. (1988). *Statistical power analysis*. Hillsdale, NJ: Erlbaum.

Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. *Cognitive Psychology*, *20*, 65-95.

Gelman, S. A., & Markman, E. (1986). Categories and induction in young children. *Cognition*, *23*, 183-209.

Gelman, S. A., & Markman, E. (1987). Young children's inductions from natural kinds: The role of categories and appearances. *Child Development*, *58*, 1532-1541.

Gutheil, G., Vera, A., & Keil, F. (1998). Do houseflies think? Patterns of induction and biological beliefs in development. *Cognition*, *66*, 33-49.

Hitch, G. J., & Towse, J. N. (1995). Working memory: What develops? In F. E. Weinert & W. Schneider (Eds.), *Memory performance and competencies* (pp. 3-21). Mahwah, NJ: Erlbaum.

Johnson S., & Solomon, G. (1997). Why dogs have puppies and cats have kittens: The role of birth in young children's understanding of biological. *Child Development*, *68*, 404-419.

Rosengren, K. S., Gelman, S. A., Kalish, C. & McCormick, M. (1991). As time goes by: Children's understanding of early growth in animals. *Child Development*, *62*, 1302-1320.

Shepp, B. E. (1978). From perceived similarity to dimensional structure: A new hypothesis about perceptual development. In E. Rosch & B. Lloyd (Eds.), *Cognition and categorization* (pp. 135-167). Hillsdale, NJ: Erlbaum.

Shepp, B. E., & Schwartz, K. B. (1976). Selective attention and the processing of integral and nonintegral dimensions: A developmental study. *Journal of Experimental child Psychology*, *22*, 73-85.

Sloutsky, V. M., & Lo, Y.-F. (1999). How much does a shared name make things similar? Part 1: Linguistic labels and the development of similarity judgement. *Developmental Psychology*, *6*, 1478-1492.

Sloutsky, V. M., & Lo, Y.-F. (2000). Linguistic labels and the development of inductive inference. In L. Gleitman & A. Joshi (Eds.), *Proceedings of the XXII Annual Conference of the Cognitive Science Society* (pp. 469-474). Mahwah, NJ: Erlbaum.

Sloutsky, V. M., Lo, Y.-F., & Fisher, A. V. (In Press). How much does a shared name make things similar? Linguistic labels and the development of inductive inference. *Child Development*.

Smith, L. (1989a). A model of perceptual classification in children and adults. *Psychological Review*, *96*(1), 125-144.

Smith, L. (1989b). From global similarities to kinds of similarities: The construction of dimensions in development. In S. Vosnuadou & A. Orthony (Eds.), *Similarity and analogical reasoning* (pp. 146-178). New York: Cambridge University Press.

Smith, L. B., & Jones, S. S. (1993). Cognition without concepts. *Cognitive Development*, *8*, 181-188.

Solomon, G. E., Johnson, S., Zaitchik, D., Carey, S. (1996). Like father, like son: Young children's understanding of how and why offspring resemble their parents. *Child Development*, *67*(1), 151-171.

Springer, K. & Keil, F. (1989). On the development of biologically specific beliefs: The case of inheritance. *Child Development*, *60*, 637-648.

Springer, K. (1996). Young children's understanding of a biological basis for parent-offspring relations. *Child Development*, *67*, 2841-2856.

Activating Verb Semantics from the Regular and Irregular Past Tense.

Catherine E Longworth (cat@csl.psychol.cam.ac.uk)
Billi Randall (bjr22@cam.ac.uk)
Lorraine K Tyler (lktyler@csl.psychol.cam.ac.uk)
Centre for Speech and Language, Dept. Exp. Psychology, Downing Site, Cambridge, CB2 3EB, UK.

William D Marslen-Wilson (william.marslen-wilson@mrc-cbu.cam.ac.uk)
MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge, CB2 2EF, UK

Abstract

Theoretical accounts of the processing of inflectional morphology make implicit, untested assumptions about the activation of verb semantics from inflected verbs. This research used semantic priming to investigate the extent to which regular and irregular past tense forms activate verb semantics, in comparison to the verb stem. The results show that past tense forms activate verb semantics to the same extent as verb stems and without differences due to verb regularity. These results provide constraining data for models of inflectional morphology.

Introduction

Research on the mental representation and processing of inflected words has focused on the type of underlying processing system required to account for our ability to produce and comprehend regularly and irregularly inflected words. The English past tense in particular has formed an important test case for such research. It provides a sharp contrast between the productive regular past tense and a limited set of irregular forms. Theoretical accounts divide into those suggesting that all inflected forms, both regular and irregular, are stored using associative memory (e.g. Rumelhart and McClelland, 1986) and those proposing that the predictability of the regular past tense allows us to compute the inflected forms from their verb stems using symbolically based rules (e.g. Pinker, 1991).

Such accounts focus primarily on the processing and representation of the regular and irregular past tense as phonological forms, tending to neglect the important issue of how these forms contact their underlying semantic and syntactic content. Nonetheless, they incorporate assumptions, explicit or implicit, about how these inflected forms map on to their semantics. Pinker has suggested that the regular past tense is not stored in its own right but computed using symbolic rules, whereas the irregular past tense is stored independently of its stem (Pinker, 1991). Comprehension of the regular past tense is assumed to require rule-based decomposition into stem and inflection. This process may delay semantic activation relative to the irregular past tense, which is not assumed to require rule-based processing.

The related account proposed by Marslen-Wilson and Tyler (1998) explicitly postulates that the regular and irregular past tense activate the semantic representations of their verb stems through different routes. The morphophonologically complex regular past tense is parsed into its stem and affix in order to access stem semantics, whereas the irregular past is recognised as a full form, which must subsequently be mapped onto verb semantics assumed to be stored at the level of the lexical entry.

Strict full-listing accounts (Butterworth, 1983) suggest that all past tense forms are stored in memory, with semantic representations presumably stored independently of their verb stems. Activating such representations would not be delayed by rule-based processing, and would not be expected to differ according to morphological complexity or verb regularity.

Parallel-distributed processing models (Joanisse and Seidenberg, 1999; Plunkett and Marchman, 1993; Rumelhart and McClelland, 1986) have assumed that both the regular and irregular past tense activate semantic units associated with their verb stems. In common with strict full-listing accounts, the fact that the regular and irregular past tense have differing degrees of phonological and orthographic similarity to their stems is not assumed to entail differential processing. However, if a mapping is less consistent (as in irregular forms) this may lead to slower or less efficient activation.

It can be seen that although these accounts focus on the consequences of the variable predictability of past tense form from its associated verb stem, they inevitably make theoretical assumptions about the activation of verb semantics from inflected forms as well. Typically one would use semantic priming to test assumptions relating to the organisation of semantic representations or the time course of its activation. However, research using semantic priming has tended to use concrete nouns and/or uninflected verbs. This is because morphologically simple words are assumed to

be the paradigmatic forms associated with semantic information, and semantic priming is thought to reflect either activation spreading between their semantic representations or a reactivation of shared properties of these representations. Morphological priming, on the other hand, uses complex words as primes, targets or both, and is a form of repetition priming. It is thought to reflect the reactivation either of an underlying morpheme (Marslen-Wilson, Tyler, Waksler and Older, 1994) or of shared semantic and form units (Plaut and Gonnerman, 2000), depending on the theoretical account. As such, the task is used to explore lexical representation and access.

What would be expected if verbs in the past tense were used as semantic primes? Ultimately they must contact verb semantics since the presence of verb inflection is a normal part of language comprehension. However, all accounts, except perhaps strict full-listing theories, assume that verbs in the past tense are processed as modified forms of their stems, as inflectional morphology does not change word meaning or form class. The stem is viewed as the most basic form of a verb and assumed to be associated with its semantic representations. Inflected verbs are assumed to be phonological modifications of the verb stem.

So how do these modified forms map on to the verb semantics associated with their stems? How long does this process of mapping take? If semantic priming is the facilitation of responses due to a semantic relationship between morphologically simple words will there be any semantic priming from the past tense at all? Alternatively there might be priming when the modified form maps onto verb semantics, but the delay due to this processing might reduce the degree of facilitation in comparison to that from the verb stem. Perhaps the greater similarity of form between the regular past tense and associated verb stems will speed up mapping onto verb semantics via the stem. This could lead to more semantic priming from the regular past than the irregular.

These questions were addressed by the following experiments using verb stems and their past tense forms as semantic, rather than morphological, primes. The motivation for this research was not to adjudicate between theories postulating one or more processing routes for inflectional morphology. Rather the aim was to provide evidence about the activation of verb semantics from inflected forms, in order to constrain currently untested assumptions implicit in these theories.

Experiment 1: Intramodal auditory semantic priming.

Morphological priming with auditory presentation of both past tense primes and verb stem targets has shown facilitation of lexical decision responses to targets following regular and irregular past tense primes (Marslen-Wilson and Tyler, 1997). It seemed prudent to begin our investigation of semantic priming from the past tense with auditory presentation of both primes and targets.

Rather than simply using regular and irregular past tense forms as semantic primes, a within-item design was selected so that each verb prime would be used in both its stem and past tense form (i.e. both "blame ACCUSE" and "blamed ACCUSE"). This allows us to establish that the uninflected forms are sufficiently semantically related to their targets to cause priming. Also, if the past tense items do prime, this can be measured against the amount of priming from the associated stem, to ascertain the effects of mapping onto verb semantics from an inflected form.

52 regular and 52 irregular verb primes were paired with semantically related verb targets. Semantic relatedness was established empirically by asking participants to rate this on a nine-point scale (1 = extremely unrelated, 9 = extremely related). Separate ratings were collected for stem and past tense forms of each prime. Groups of 15 participants (native speakers of UK English, aged between 18 and 40, with no language disorders) rated the semantic relatedness of either the stem or past tense form of each verb prime paired with its target. Past tense prime-target pairs were rated slightly less related than their associated stem prime-target pairs (mean rating for stems = 7.34, sd .63, mean rating for the past tense = 7.12 sd .73, $F(1,96) = 13.54$, $p < .001$) with no effect of verb regularity.

Unrelated primes (e.g. "laugh ACCUSE" and "laughed ACCUSE") were selected by rotating test primes about their targets whilst maintaining tense and verb regularity. This ensured that there could be no systematic differences between test and control primes in each condition, other than semantic relatedness to targets. The semantic relatedness of control primes and their targets was pretested in the same way as the test primes. Past tense prime-target pairs were again rated as slightly less related than their associated stem prime-target pairs (mean rating for stem controls = 2.39, sd .84, mean rating for past tense controls = 2.07 sd .65, $F(1,96) = 13.16$, $p < .001$) with no effect of verb regularity. Related test primes had a mean rating of 7.23 (sd = .69) and unrelated controls had a mean rating of 2.2 (sd = .75). As our aim was to examine semantic rather than associative priming, all primes were selected to have a low associative strength to their targets.

To ensure that any differences in semantic priming between regular and irregular past tense primes were due to verb regularity we matched primes across verb regularity for semantic relatedness and associative strength to targets, familiarity, imageability, and surface and cumulative frequencies from the Celex Lexical

Database (Baayen, Piepenbroek and Gulikers, 1995). Number of syllables could not be matched, as the regular past tense tends to be longer than the irregular form. Targets were matched across conditions for surface and cumulative frequencies, familiarity, imageability and number of syllables. Since many English verbs can also be used as nouns, we ensured that all primes had higher surface and cumulative frequencies as verbs.

The listener's task was to make a lexical decision to each target, with instructions to respond as quickly and accurately as possible. A range of fillers was selected to ensure that semantic relationships, verbal primes or inflected verb primes could not be used to predict word targets. To this end we used 208 unrelated noun/adjective prime-target pairs, with half the nouns in the plural form and nonwords used for half the targets. We also used 104 verb-nonword pairs with the same proportions of regular and irregular, stem and past tense primes as the test items.

The materials were divided into four versions of the experiment. These were balanced so that all targets appeared once in each version. Each version had the same target preceded by either a semantically related stem or past tense or a control stem or past tense. All versions had 460 trials: 24 practice trials, 20 "warm-up" trials, 52 test trials (13 of each condition), 52 control trials (13 in each condition) and 312 filler trials. Semantically related verb pairs made up 25% of the word targets heard. These were pseudo-randomly distributed throughout the list, with the same order of test and filler items in each of the four versions. There were an equal number of word and nonword targets in each version.

All items were recorded by a female native speaker of English onto DAT tape. They were digitized at a sampling rate of 22kHz, and were played binaurally to the listeners over headphones under the control of DMDX experimental software (Forster and Forster, 1990).

There was a 200 millisecond interval between primes and targets and participants had up to 3 seconds to respond. After responding the next trial followed in 1500 milliseconds. Reaction times were measured from target onset. The experiment lasted approximately 50 minutes in total.

60 participants (native speakers of UK English, aged between 18 and 40, with no language disorders) took part in the experiment. 15 participants were randomly assigned to each version of the experiment.

Results

The data from four participants were discarded because of relatively high error rates and unusual, or variable reaction times. Six items were also removed, four because of experimenter error and two because of high error rates. This left a total of 56 participants and 98 items.

For the analysis of reaction times, all errors (2.5%) and extreme values (0.1%, defined as $\leq 500 \geq 2000$ msec) were removed from the data. Mean reaction times were then calculated over participants and items. These were entered into two analyses of variance on participant (F_1) and item (F_2) means, with the factors of prime type (test or control), verb regularity (regular or irregular), tense (stem or past tense) and version (1-4). Item means are shown in Table 1.

There was a main effect of prime type (F_1 (1,52) = 166.53, $p < .001$; F_2 (1,90) = 91.31, $p < .001$) due to faster reaction times following semantically related (mean RT = 826 msec, sd = 82 msec) compared to unrelated (mean RT = 875 msec, sd = 89 msec) primes.

There was also a main effect of tense (F_1 (1,52) = 14.37, $p < .001$; F_2 (1,90) = 15.69, $p < .001$) with reaction times following past tense items (mean RT = 857 msec, sd = 87 msec) being slower than those following verb stems (mean RT = 843 msec, sd = 90 msec). There was no main effect of verb regularity. There were no interactions between priming, tense and regularity. Planned comparisons confirmed that there was significant priming for every condition.

Discussion

This first experiment found that past tense primes significantly facilitated lexical decision responses to semantically related targets. With auditory presentation these inflected words were able to map onto their verb semantics sufficiently strongly and quickly to prime responses to related words presented 200 milliseconds later. Not only did the related past tense items facilitate responses but the main effect of priming and the absence of any interaction between priming and tense shows that they primed as much as their associated stems. This is surprising as semantic priming is

Table 1 Mean item reaction times and standard deviations for intramodal semantic priming.

	STEM PRIMES			PAST TENSE PRIMES		
	Test	Control	Diff	Test	Control	Diff
REGULAR	818 (82)	872 (85)	54 ***	840 (74)	883 (82)	43 ***
IRREGULAR	815 (86)	868 (94)	53 ***	830 (85)	876 (95)	46 ***

*** $p < .001$

generally thought to reflect semantic relationships between basic lexical forms. As the past tense is an inflected, or modified, form of its verb stem one might not have expected both to prime equally, especially as past tense prime-target pairs had been rated as being less related than their associated stem prime-target pairs in pretests.

The phonological form of the regular past tense is nearly identical to that of its stem so perhaps it is more predictable that this could map onto verb semantics as quickly as its stem. However, the phonology of words in the irregular past tense is less similar to their verb stems than the regular past tense, so we might have expected these to prime less. The lack of a significant interaction between priming, tense and regularity shows that this was not the case. Both the regular and irregular past tense primed as much as their stems and as much as each other.

There are several possible interpretations of these findings. If inflected verbs need to map on to the stem to activate verb semantics this mapping may occur so quickly and effectively that it does not interfere with priming. The degree of phonological similarity to the stem does not seem to affect the efficiency of this mapping. The irregular past tense has less phonological similarity to associated stems than the regular past tense yet this does not delay access to verb semantics. If the regular and irregular past tense are processed differently this does not seem to have consequences for semantic activation as measured in this experiment.

Another alternative is that inflected verbs do not need to map on to the stem in order to activate verb semantics. All forms might be equally associated with verb semantics. However, if the stem is not the most basic form associated with verb semantics, why should past tense inflections have led to slower responses? It is not the case that there was no effect of words being inflected in this experiment, just that this did not interact with semantic priming.

To summarize, this experiment found that past tense inflection, whether regular or irregular, did not affect semantic priming, although it did increase response latencies to targets.

Experiment 2: Cross-modal semantic priming.

It is already established that intramodal presentation of the regular and irregular past tense facilitates responses to associated verb stems (Marslen-Wilson and Tyler, 1997) and Experiment 1 established that they also facilitate responses to semantically related verbs. However, when primes and targets are presented in different modalities the irregular past tense no longer facilitates responses to associated stems (Marslen-Wilson, Hare and Older, 1993). This contrasts with the regular past tense, which continues to show morphological priming. As modality seems to affect morphological priming from the irregular past tense the second experiment tested whether it also affects semantic priming.

Cross-modal presentation also allows us to probe semantic activation earlier in its time course. The first experiment might have failed to detect effects of tense or verb regularity on semantic activation if these had dissipated in the 200 millisecond inter-stimulus interval between the prime and target. In addition, reducing available processing time allows us to assess whether the tense effect found in the first experiment reflected the impact of inflected words on initial semantic activation or a possible post-lexical effect arising in the interval between hearing primes and responding to targets.

The second experiment used the design and materials from Experiment 1 but this time a visual target presented for 500 milliseconds followed the immediate offset of the auditory prime. Reaction times were recorded from target onset as before. 51 participants took part (native speakers of UK English, aged between 18 and 40, with no language disorders). 12 to 14 participants were randomly assigned to each version of the experiment.

Results

The data from four participants were discarded due to unusual and/or variable reaction times. The items involving 5 target words were also removed because of high error rates. This left a total of 47 participants and 99 test items.

For the analysis of reaction times, all errors (2%) and extreme values (0.1%, defined as ≥ 1350 msec) were removed from the data. Mean reaction times were then calculated over participants and items. These were

Table 2 Mean item reaction times and standard deviations for cross-modal semantic priming.

	STEM PRIMES			PAST TENSE PRIMES		
	Test	Control	Diff	Test	Control	Diff
REGULAR	516 (51)	525 (54)	9	511 (46)	538 (63)	27 ***
IRREGULAR	517 (50)	536 (60)	19 **	512 (44)	531 (59)	19 *

* p < .05, ** p < .01, *** p < .001

entered into two analyses of variance on participant (F_1) and item (F_2) means, with the same factors as experiment 1. Item means are shown in Table 2.

There was a main effect of prime type (F_1 (1,43) = 28.62, $p < .001$; F_2 (1,91) = 24.59, $p < .001$), due to faster reaction times following related (mean RT = 514 msec, sd = 48 msec) compared to unrelated (mean RT = 532 msec, sd = 59 msec) primes. There were no main effects of verb tense or regularity and no interactions between priming, tense and regularity.

Planned comparisons on item means in individual conditions indicated that regular stems did not prime significantly (t (48) = -1.23, p = .225) due to an unexplained interaction between priming and version (F_2 (3,45) = 6.49, $p <.001$). One version showed reduced latencies following semantically unrelated regular stems. The remaining three versions showed a significant effect of priming (F_2 (1,33) = 9.62, p = .004) and no interaction with version (F_2 (2,33) = 2.39, p = .107) with semantically related regular stems (RT = 510 msec) facilitating responses by 21 msec relative to unrelated regular stems (RT = 531 msec). All other conditions produced significant priming across all four versions.

Discussion

This experiment confirmed the main effect of priming, with semantically related primes facilitating responses to targets as before. Priming did not interact with tense or verb regularity. The amount of priming shown by regular stems is smaller than the other conditions when all four versions are analyzed. The main effect of priming, however, indicates that regular stem priming is not significantly different to other conditions when version-related variance is partialled out.

Past tense primes facilitated semantically related targets despite cross-modal presentation and reduced processing time. As the irregular past tense fails to act as a morphological prime under these conditions, one might have expected to see an interaction between prime, tense and verb regularity, such that the irregular past tense failed to prime despite priming in all other conditions. This was not found. Again there seem to be no consequences for semantic activation, as indexed by semantic priming, of words being inflected or having different degrees of phonological similarity to their stems.

In addition, the lack of a tense effect, when processing time is reduced, suggests that the effect found in the earlier experiment did not reflect the impact of inflected words on initial semantic activation but a post-lexical effect arising in the interval between hearing primes and responding to targets. This might be a consequence of the irrelevance of the past tense inflection to the subsequent stem target.

Thus in this second experiment, once processing time was reduced, there was no effect of words having past tense inflections on semantic activation. This suggests that all forms of a verb access its semantic representations equally rapidly.

General discussion.

Research on the English past tense has concentrated on issues relating to phonological or orthographic form. A central question has been whether the predictable similarity in form between verb stems and the regular past tense engages specialized rule-based processes to compute the past tense rather than storing it in full. However, theoretical models answering this question have made implicit, largely untested assumptions about the activation of verb semantics from verbs in the past tense.

Most models assume that verbs in the past tense access the same semantic representations as their verb stems and are processed as modified forms of their verb stems. We assume that the form, not the semantics, of the verb is modified as a result of syntactic constraints. These assumptions lead us to expect some effect of this modification of form on the activation of verb semantics. Single route accounts suggest that inflected words will have patterns of phonological activation that are highly similar, but not identical, to those of their stems. Dual route accounts suggest that inflected words must map onto underlying morphemic representations, accessed via the verb stem. Therefore, we might expect not to see semantic priming from inflected forms, or to see reduced priming reflecting the time taken to map onto semantics via the verb stem. However, both the experiments reported here show that this is not the case. Activation of verb semantics did not show any effects of verb inflection. The only consequence of verbs being inflected, i.e. the increase in response latencies to auditory targets following past tense primes, did not interact with semantic priming and appeared to be a post-lexical integration effect, as it was not present when processing time was reduced in the second experiment.

It is also commonly assumed that the amount and predictability of phonological similarity between stems and past tense, which is greater in the regular past tense, will have processing consequences, and might even engage different types of processing. The results reported here, however, suggest that these factors have no effect on the activation of verb semantics in comprehension.

Thus both stem and past tense, regular and irregular verb primes all accessed verb semantics equally in the same time frame. Activation of verb semantics seems insensitive to morphological complexity and inflectional regularity. The surprising aspect of this is

that verb stems are assumed to be the most basic form of verbs and processing of the regular and irregular past tense has been found to dissociate in development (Berko, 1958) and to doubly dissociate following neurological damage (Marslen-Wilson and Tyler, 1997; Tyler, deMornay-Davis, Anokinah, Longworth, Randall, & Marslen-Wilson, in press; Ullman, Corkin, Coppola, Hicock, Growdon, Koroshetz, and Pinker, 1997).

In particular these results might seem surprising given the dissociations between processing the regular and irregular past tense even in the normal adult. Both the regular and irregular past tense act as morphological primes when presented intramodally but with cross-modal presentation the irregular past tense no longer primes stem targets (Marslen-Wilson, Hare and Older, 1993). However, the current results show that the irregular past tense does facilitate responses to semantically related verbs, both cross-modally and intra-modally. This is consistent with the complete equivalence of regular and irregular forms in terms of their linguistic and communicative function.

In summary, these experiments suggest that past tense forms activate the same semantic representations as their stems, priming related words to the same extent. There is no evidence that morphophonological processing delays semantic activation. There was no reduction in semantic priming to suggest a processing cost for inflected verbs and no regularity differences to suggest that degree of phonological modification affects the time course of access to semantics. This suggests a lexical architecture permitting either direct mapping of all verb forms onto semantics, or the mapping of all verb forms, regular and irregular, stem and past tense, onto an abstract root morpheme providing access to semantics. If, on either view, separate processing routes are indeed involved in the perceptual analysis of regular and irregular forms, then they deliver their output to higher-order interpretive systems with essentially the same time-course. Theoretical models explaining the processing of the regular and irregular past tense therefore need to bear in mind that access to verb semantics seems to be insensitive to both morphological complexity and inflectional regularity.

Acknowledgements

This research was supported by an MRC post-graduate research studentship (number G78/6297), and an MRC programme grant to LKT.

References

Baayen, R. H., Piepenbroek, R. and Gulikers, L. (1995) The CELEX Lexical Database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Berko, J. (1958) The child's learning of English morphology. *Word, 140*, 150-177.

Butterworth B. (1983) Lexical Representation. In B. Butterworth (ed.) *Language Production volume 2: Development, writing and other language processes.* Academic Press.

Coltheart, M. (1981) MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, *33A*, 497-505.

Forster, K.I., and Forster, J.C. (1990) The DMASTR display system for mental chronometry. Tucson, Arizona: University of Arizona.

Joanisse, M.F., and Seidenberg, M.S. (1999) Impairments in Verb Morphology Following Brain Injury: A Connectionist Model. *Proceedings of the National Academy of Sciences, USA. 96*,7592-7597.

Marslen-Wilson, W.D., Hare, M. and Older, L. (1993). Inflectional morphology and phonological regularity in the English mental lexicon. In Proceedings of the 15th Annual Conference of the Cognitive Science Society. London.

Marslen-Wilson, W.D., Tyler, L.K., Waksler, R. and Older, L. (1994) Morphology and meaning in the English mental lexicon. *Psych. Rev. 101 (1),* 3-33.

Marslen-Wilson, W.D., and Tyler, L.K. (1997) Dissociating types of mental computation. *Nature, 387,*592-594.

Marslen-Wilson, W.D., and Tyler, L.K. (1998) Rules, representations and the English past tense. *Trends in Cognitive Sciences, 2 (11),*428-435.

Pinker, S. (1991) Rules of Language. *Science, 253,* 530-535.

Plaut, D.C., and Gonnerman, L.M. (2000) Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Lang. Cognitive Proc. 15 (4-5),* 445-485.

Plunkett, K., and Marchmann, V. (1993) From rote learning to system building: acquiring verb morphology in children and connectionist nets. *Cognition, 48,* 21-69.

Rumelhart, D., and McClelland, J. (1986) On learning the past tense of English verbs. In McClelland, J.L., Rumelhart, D.E., and the PDP Research Group, (Eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol.2. Psychological and Biological Models*), MIT Press.

Tyler, L.K., deMornay-Davies, P., Anokhina, R.A., Longworth, C.E., Randall, B., & Marslen-Wilson, W. D. (in press) Dissociations in processing past tense morphology: Neuropathology and behavioral studies. *Journal of Cognitive Neuroscience*.

Ullman, M.T., Corkin S., Coppola M., Hicock, G., Growdon, J.H., Koroshetz, W.J., and Pinker, S. (1997) A neural dissociation within language: evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of Cognitive Neuroscience, 9,* 266-276.

Towards a Theory of Semantic Space

Will Lowe (wlowe02@tufts.edu)
Center for Cognitive Studies
Tufts University; MA 21015 USA

Abstract

This paper adds some theory to the growing literature of semantic space models. We motivate semantic space models from the perspective of distributional linguistics and show how an explicit mathematical formulation can provide a better understanding of existing models and suggest changes and improvements. In addition to providing a theoretical framework for current models, we consider the implications of statistical aspects of language data that have not been addressed in the psychological modeling literature. Statistical approaches to language must deal principally with count data, and this data will typically have a highly skewed frequency distribution due to Zipf's law. We consider the consequences of these facts for the construction of semantic space models, and present methods for removing frequency biases from semantic space models.

Introduction

There is a growing literature on the empirical adequacy of semantic space models across a wide range of subject domains (Burgess et al., 1998; Landauer et al., 1998; Foltz et al., 1998; McDonald and Lowe, 1998; Lowe and McDonald, 2000). However, semantic space models are typically structured and parameterized differently by each researcher. Levy and Bullinaria (2000) have explored the implications of parameter changes empirically by running multiple simulations, but there has up until now been no work that places semantic space models in an overarching theoretical framework; consequently there there are few statements of how semantic spaces *ought* to be structured in the light of their intended purpose.

In this paper we attempt to develop a theoretical framework for semantic space models by synthesizing theoretical analyses from vector space information retrieval and categorical data analysis with new basic research.

The structure of the paper is as follows. The next section brie¤y motivates semantic space models using ideas from distributional linguistics. We then review Zipf's law and its consequences the distributional character of linguistic data. The £nal section presents a formal de£nition of semantic space models and considers what effects different choices of component have on the resulting models.

Motivating Semantic Space

Firth (1968) observed that "you shall know a word by the company it keeps". If we interpret company as *lexical* company, the words that occur near to it in text or speech, then two related claims are possible. The £rst is unexceptional: we come to know about the syntactic character of a word by examining the other words that may and may not occur around it in text. Syntactic theory then postulates latent variables e.g. parts of speech and branching structure, that control the distributional properties of words and restrictions on their contexts of occurrence. The second claim is that we come to know about the *semantic* character of a word by examining the other words that may and may not occur around it in text.

The intuition for this distributional characterization of semantics is that *whatever* makes words similar or dissimilar in meaning, it must show up distributionally, in the lexical company of the word. Otherwise the supposedly semantic difference is not available to hearers and it is not easy to see how it may be learned.

If words are similar to the extent that they occur in the similar contexts then we may de£ne a statistical replacement test (Finch, 1993) which tests the meaningfulness of the result of switching one word for another in a sentence. When a corpus of meaningful sentences is available the test may be reversed (Lowe, 2000a), and under a suitable representation of lexical context, we may hold each word constant and estimate its typical surrounding context. A semantic space model is a way of representing similarity of typical context in a Euclidean space with axes determined by local word co-occurrence counts. Counting the co-occurrence of a target word with a £xed set of D other words makes it possible to position the target in a space of dimension D. A target's position with respect to other words then expresses similarity of lexical context. Since the basic notion from distributional linguistics is 'intersubstitutability in context', a semantic space model is effective to the extent it realizes this idea accurately.

Zipf's Law

The frequency of a word is (approximately) proportional to the reciprocal of its rank in a frequency list (Zipf, 1949; Mandelbrot, 1954). This is Zipf's Law. Zipf's law ensures dramatically skewed distributions for almost

all statistics applied to language; the power scaling ensures that the majority of words occur very infrequently, creating a severe sparse data problem, and that the top few most frequent words constitute the majority of all tokens. For example, the 10 most frequent word stems, or lemmas, in the 100M word British National Corpus are 'the', 'be', 'of', 'and', 'to', 'a', 'in', 'have', 'that' and 'it', constituting slightly over one quarter of all tokens in the corpus (25974687 / 99985962 0.26). Also the most frequent words of English are grammatical functors or closed class words (Cann, 1996), which although vital to syntax, are typically uninformative with respect to word meaning. Much of the next sections will be devoted to dealing with the distributional effects of Zipf's law.

To introduce some notation, semantic space models typically represent the distributional context of each word t in terms of a set of representative 'context' words $b_1 \ldots b_D$. t's distributional pro£le is then represented by a vector of co-occurrences $\mathbf{v}$ where $\mathbf{v}_i$ is a function of $f^W(b_i,t)$, the number of times b_i occurs in a window W words either side of t in a corpus of N words. For future reference $f(t)$ is the occurrence frequency of t in the corpus, $p(t)$ is the probability of t, often estimated by t N, and $p^W(b_i,t)$ is the probability of seeing b_i and t together in a window of size W.

Semantic Space

A semantic space model is method of assigning each word in a language to a point in a real £nite dimensional vector space. Formally it is a quadruple $\mathtt{A},\mathtt{B},\mathtt{S},\mathtt{M}$:

$\mathtt{B}$ is a set $b_{1\ldots D}$ of basis elements that determine the dimensionality D of the space and the interpretation of each dimension. $\mathtt{B}$ is often a set of words (Lund et al., 1995, e.g.) although lemmas (Lowe and McDonald, 2000), encyclopedia articles (Landauer and Dumais, 1997) and whole documents have been used.

$\mathtt{A}$ speci£es the functional form of the mapping from co-occurrence frequencies between particular basis elements and each word in the language so that each word is represented by a vector $\mathbf{v} = \mathtt{A}(b_1,t), \mathtt{A}(b_2,t), \ldots, \mathtt{A}(b_D,t)$. $\mathtt{A}$ may be the identity function.

$\mathtt{S}$ is a similarity measure that maps pairs of vectors onto a continuous valued quantity that represents contextual similarity.

$\mathtt{M}$, is a transformation that takes one semantic space and maps it onto another, for example by reducing its dimensionality. Various choices for these elements are possible, and lead to rather different spaces. $\mathtt{M}$ may also be an 'identity' mapping that does not change the space. In the following sections we consider the implications of different choices of $\mathtt{A},\mathtt{B},\mathtt{S}$ and $\mathtt{M}$.

$\mathtt{A}$: Lexical Association Function

Zipf's law suggest that using vectors of co-occurrence counts directly may not be a good choice when constructing a semantic space. To see why, consider two words t_1 and b with probabilities $p(t_1)$ and $p(b)$. If t_1 and b have *no* semantic relation to each other, then they will be distributionally related to one another only through their syntactic properties e.g. by the fact that they are both nouns. For simplicity we ignore any residual syntactic dependence and model their empirical frequencies $f(t_1)$ and $f(b)$ as independent binomially distributed random variables

$$\begin{array}{ll} f(t_1) & B(p(t_1),N) \\ f(b) & B(p(b),N). \end{array}$$

In this idealization t_1 and b are perfectly distributionally independent so $f^W(b,t_i) = WN\,p(b,t_1) = WN\,p(t_1)p(b)$ (this is just the expected co-occurrence frequency summed over each possible position in the window).

The fact that the expected co-occurrence count under independence is linear in the probability of t_1 leads to a problem in any model that sets $\mathtt{A}((b,t_i) = f^W(b,t_i)$, e.g. the Hyperspace Analogue to Language (HAL; Lund et al., 1995). Even if t_1 and t_2 are unrelated, if $p(t_1)$ $p(t_2)$ then their vectors will contain elements with similar magnitudes. This implies that any similarity measure applied to the vectors will judge them to be similar. Conversely if they are related but $p(t_1)$ $p(t_2)$ then their vectors will contain elements with widely differing magnitudes, simply due to their differing occurrence probability. Zipf's Law threatens that any difference in distributional pro£le available in $f^W(b,t_i)$ may be swamped by the effect of a difference in occurrence probability.

The upshot for models such as the HAL that use vectors of counts that are not corrected for chance is that distances will have a frequency bias. That is, proximity on semantic space will be partly due to distributional similarity, and partly due to relative frequency; the larger the difference in occurrence probability, the larger association a context element must have to affect the similarity function.

Since it is unlikely that semantic similarity depends on relative frequency, we have a theoretical reason not to use raw co-occurrence counts as a lexical association function.

Researchers in information retrieval have also noted problems with raw co-occurrence counts and use various weighting schemes to counteract them. Latent Semantic Analysis (LSA; Landauer and Dumais, 1997; Rehder et al., 1997), a semantic space model derived from information retrieval research uses an entropy-weighted function: $\mathtt{A}(b,t) \propto \log(f^W(b,t)$ 1). The logged co-occurrence count is then divided by the entropy of the distribution of b over each documents. If b is evenly distributed across documents then it is probably not informative about any particular document. In contrast if it occurs in some but not others it may be more informative about their content.

LSA's lexical association function is designed to allow arbitrarily many basis elements into the similarity calculation by weighting them appropriately. However neither logging nor dividing by entropy is guaranteed to reverse the effects of chance co-occurrence since this is never explicitly estimated.

	Target	Non-target
Context	$f^W(b,t)$	$f^W(b,\ t)$
Non-context	$f^W(\ b,t)$	$f^W(\ b,\ t)$

Table 1: Co-occurrence frequency within a window of target, context and all other words. t represents a word that is not t.

Lowe and McDonald (2000) used a log-odds-ratio measure to explicitly factor out chance co-occurrences. The empirical counts necessary for computing the log-odds-ratio are shown in Table 1. t represents any word that is not t, b represents a word that is not the context word b and $f^W(\ b,t)$ is the number of times a word that is not the context word occurs among the W words surrounding t.

Computing the cell counts is straightforward because there exists a very close approximation that is a function only of $f^W(b,t)$ itself, $f(t)$, $f(b)$, W, and N:

$$\begin{aligned} f^W(b,\ t) &= Wf(b) \quad f^W(b,t) \\ f^W(\ b,t) &= Wf(t) \quad f^W(b,t) \\ f^W(\ b,\ t) &= WN \quad (f^W(b,\ t) \quad f^W(\ b,t) \\ &\qquad f^W(b,t)). \end{aligned}$$

To derive these expressions consider the limiting situation where $W = 1$ and $f(b,t)$ is the number of times the bigram b,t occurs. Since by de£nition $f(b) = f(b,t) \quad f(b,\ t)$, then $f(b,\ t) = f(b) \quad f(b,t)$, and the same reasoning applies to $f(\ b,t)$. Similarly the number of elements in the table, $f(b) \quad f(\ b)$, must be the number of bigrams in the corpus. For a large corpus this is essentially N, the number of words in the corpus. Therefore since $f(\ b,\ t)$ is the only cell undetermined it is obtained by subtracting the sum of the other cells from N. The W factors appear on quantities other than the co-occurrence count when the window size is more than one because only $f^W(\ b,\ t)$ already takes the window size into account[1].

We obtain probabilities from Table 1 by dividing each cell count by WN. Then the *odds* of seeing t rather than some other word when b is present are $p^W(b,t)\ \ p^W(b,\ t)$, and the odds of seeing t in the absence of b is $p^W(\ b,t)\ \ p^W(\ b,\ t)$. Therefore if the presence of b *increases* the probability of seeing t then the odds ratio (Agresti, 1990)

$$\begin{aligned} \theta(b,t) &= \frac{p^W(b,t)\ \ p^W(b,\ t)}{p^W(\ b,t)\ \ p^W(\ b,\ t)} \\ &= \frac{p^W(b,t)\,p^W(\ b,\ t)}{p^W(b,\ t)\,p^W(\ b,t)} \end{aligned}$$

[1]The derivation is reported elsewhere (Lowe, 2000a).

is greater than 1. When the presence of b makes no difference to the probability of seeing t then $\theta = 1$ and we can conclude that b and t are distributionally independent. Finally, if $\theta < 1$ the presence of t makes seeing b less probable.

We can estimate the odds ratio from Table 1:

$$\hat{\theta}(b,t) = \frac{f^W(b,t)\,f^W(\ b,\ t)}{f^W(b,\ t)\,f^W(\ b,t)}.$$

Where the WN factors have canceled. This measure is often logged so that then the magnitude of $\log\hat{\theta}(b,t)$ can be interpreted as a direct measure of the level of associative strength between t and b, with the effects of chance co-occurrence factored out. Positive values indicate greater than chance positive association.

Lexical Association in Lexicography

The *most* informative words for t are those that occur only in its context, e.g. t='sealed' and b='hermetically'. Instances of word pairs like this are concordances, or collocations, and are of interest to lexicographers. Consequently, the log-odds-ratio also provides a method of £nding collocations between words. Previous work in lexicography has used pointwise mutual information, log-likelihood ratios, and T-tests. Since by symmetry these alternative measures can also be lexical association functions, we review them brie¤y below.

Mutual Information The pointwise mutual information $I(b,t)$ between t and b Church and Hanks (1990) is

$$I(b,t) = \log\frac{p^W(b,t)}{Wp(b)p(t)}$$

and can be also be estimated using the frequencies in Table 1. $I(b,t)$ measures how much information an occurrence of b contains about t. If b occurs with t no more often than would be expected by chance then $p^W(b,t) = Wp(b)p(t)$ and $I(b,t) = 0$, so the mutual information measure effectively factors out random co-occurrences. However, if t and b always occur together then $p^W(b,t) = p(b)$ and $I(b,t) = \log 1 \ \ p(t)$, so the less frequent b and t are the larger their association is. In contrast, changing the marginal probabilities of t or b is equivalent to adding a constant value to rows or columns of the contingency tables above (Bishop et al., 1975). It is easy to con£rm that this change makes no difference to θ.

The G-score Dunning (1993) uses a log-likelihood ratio statistic (Agresti, 1990), which he calls the G-score, to discover collocations in text. This method compares two models of the relationship between t and b. In the £rst model (association) assumes that $p(b \quad t) = p(b \quad t)$, whereas the second model (no association) assumes that $p(b \quad t) = p(b \quad\ \ t)$. The statistic is the ratio of the maximized log-likelihoods for each model's parameters. This measure takes chance co-occurrence into account because it implicitly compares the observed co-

occurrence frequencies with the co-occurrence frequencies that would be expected by chance. For example, the expected value of the top left cell in Table 1 is $Wf(t)f(b)\ N$ under (no association) but $f^W(b,t)$ under (association). Empirically using log-likelihood ratios as vector elements in a semantic space generates similar results to using log-odds-ratios. This is to be expected since both measures take chance co-occurrences into account. Alternative measures include the χ^2 statistic and Fisher's exact test. However, Dunning shows that the distributional properties of the G-score are superior under normal lexicographic conditions, and the hypergeometric probabilities required in Fisher's test are intractable to compute for contingency tables containing very large counts (Agresti, 1990). For example, $f^W(\ b\ \ t)$ will typically exceed the number of words in the corpus.

Considering the lexicographic task emphasizes the 'second order' nature of semantic space measures of similarity: they re¤ect regularities across multiple '£rst order' association measures, one for each vector element. This interpretation is taken up again in discussing appropriate similarity functions below.

B : Choosing a Basis

When choosing basis elements for a semantic space there is a trade-off between choosing words that are representative of sentence content, but may not give reliable count statistics due to their low frequency, and choosing high frequency words that provide reliable statistics but appear in almost every sentence of the language. The trade-off is an instance of the bias-variance dilemma in statistical learning theory (Geman et al., 1992).

The Bias-Variance Dilemma Every statistical model is able to represent a subset of the class of possible hypotheses about data. The range of hypotheses is typically controlled by the model's structure and by a set of adjustable parameters. More ¤exible models can represent more hypotheses and are said to have less *bias*. In contrast, a very ¤exible model will require a large amount of data to determine accurate values for its parameters. When there is not enough data compared to the number of parameters, parameter estimates may be optimal for the particular data set the model was trained on, but will fail to generalize to new data. A model that 'over£ts' in this way is said to have high *variance*. Model variance can be decreased at the cost of adding bias e.g. by constraining or removing parameters. Bias can be decreased by making the model more ¤exible, at the cost of needing more data to cope with increased variance.

In a semantic space the vector elements, $\mathtt{A}(b,t)$ are parameters that estimate the amount of association between b and t on the basis of observed data $f^W(b,t)$. When choosing the basis elements $b_1 \ldots b_D$, we can de£ne a highly biased model by choosing only very high frequency words. Co-occurrence counts for high frequency words are very reliable because high frequency words appear in nearly all sentences. This biased model will have very low variance; each $\mathtt{A}(b,t)$ is a well-determined parameter because $f^W(b,t)$ is large enough to provide a reliable estimate of $p^W(b,t)$. However, every vector will be similar because all words in the language tend to occur with the high frequency words in the basis, irrespective of their distributional pro£le. Consequently, distances between words will be extremely similar and vectors in the biased model will fail to re¤ect important distributional differences.

Alternatively, if only low frequency content words are chosen as basis elements then vectors will be more highly informative and distances in the space will be able to re¤ect subtle distributional similarities. This model will have high variance because the co-occurrence counts needed to determine $\mathtt{A}(b,t)$ are unreliable. Variance can always be decreased by providing more data, but Zipf's law suggests a power relation between the amount of new text that would need to be found and the reduction in co-occurrence count variability.

In theory the fullest possible distributional pro£le for a word would include all words in the language, generating an infeasibly large vector. In practice this is not possible and some subset of words must be chosen.

The solution for LSA is to use as many words as possible with appropriate weighting for each vector element, and then use **M** to compress the original vectors into a smaller space with dimensions that are linear combinations of the original ones.

The Column Variance Method For HAL, elements of B are chosen by compiling a 70,000 70,000 matrix of word co-occurrences and discarding the columns of lowest variance[2]. Consistent with Zipf's law, column variance decreases sharply with the frequency of the word corresponding to the column (Lund et al., 1995). Then for each set of experimental stimuli, Burgess *et al.* compute variances over each vector element and retain only the most variant. We can refer to this as the column variance method of basis element choice.

The method is dif£cult to analyze because the basis is recomputed for each experiment, but we can show that it has a frequency bias. If b and t are unrelated then we can, again, model them as Binomially distributed. In the simple case where $W = 1$, the variance of the frequency count under independence is

$$\begin{aligned} \mathrm{Var}\, f^W(b,t) &= Np(t)p(b)(1 \quad p(t)p(b)) \\ &= Np(t)p(b) \quad Np(t)^2p(b)^2. \end{aligned}$$

so the expected variance of $f^W(b,t)$ is quadratic in $p(b)$. The expected variance of the elements of a *column* of such counts is the same as the variance of the column sum i.e. the sum of the individual variances. Figure 1 shows the expected variances for a 14 14 table of co-occurrence counts for perfectly unrelated words with occurrence probabilities ranging from 0.5 to 0.0667. Even completely unrelated words will show distinct structure

[2]Co-occurrences are also weighted by distance, but this does not affect the following argument.

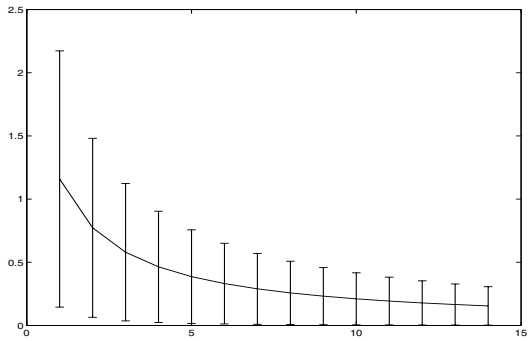


Figure 1: An example of column variance method. Expected column means based on expected co-occurrence counts between each of 14 hypothetical unrelated words. To estimate means and variances for a corpus of N words, multiply all quantities by N. Error bars represent expected column variances.

in their column variances, but this is entirely due to their baseline frequencies.

There are two possible causes for a high column variance. The £rst cause is simple frequency as shown in Figure 1. The second reason is that the words are in fact distributionally related. Then unexpectedly large variance can be a sign that the Binomial assumption has failed, and that two words are in fact related. However the size of the variance increase necessary is variable. In the column variance method, for a word that is distributionally related to some of the experimental materials to make it into the £nal basis set it must be strongly associated enough that its observed column variance moves it into the window of very high variance words at the upper end of the frequency table. In other words, it is not enough to be twice as variant as would be expected by chance, a word must be as many times more variant as it takes to have a variance that is absolutely high; lower frequency words have to work harder and unrelated but high frequency words will get chosen anyway.

This analysis of the column variance method predicts that, in the absence of strong association, the variance of a column corresponding to some candidate element will correlate strongly with that element's frequency.

This was tested by taking candidate lemmas of frequency rank 100 to 600 in the BNC, and experimental stimuli from McKoon and Ratcliff's graded priming study (see Lowe and McDonald, 2000). The analysis predicts that the levels of genuine association (corrected for frequency) between these candidates and the experimental stimuli will be be low because the words are so frequent that they provide little information about context. In fact for this data log-odds-ratios are mildly *negatively* correlated with column variance r=-.317 p<.001. In contrast candidate frequencies strongly positively correlated with column variance for co-occurrence counts, r=.8553 p<.001.

S : Similarity Measure

Two popular similarity measures are Euclidean distance and the cosine. For two vectors $\mathbf{v}$ and $\mathbf{w}$ in a D-dimensional basis, the squared Euclidean distance $\|\mathbf{v} - \mathbf{w}\|^2$ is simply related to the cosine $\rho_{\mathbf{vw}}$ of the angle between them:

$$\begin{aligned} \|\mathbf{v} - \mathbf{w}\|^2 &= \sum_{i=1}^{D} (v_i - w_i)^2 \\ &= \|\mathbf{v}\|^2 \; \|\mathbf{w}\|^2 \; 2\frac{\mathbf{vw}}{\|\mathbf{v}\| \; \|\mathbf{w}\|} \\ &= \|\mathbf{v}\|^2 \; \|\mathbf{w}\|^2 \; 2\rho_{\mathbf{vw}} \end{aligned}$$

where $\|\mathbf{w}\|^2 = \sum_i^D w_i^2$ is a squared vector length. From this equation it can be seen that $\|\mathbf{v} - \mathbf{w}\|^2 \propto \rho_{\mathbf{vw}}$ only when $\mathbf{v}$ and $\mathbf{w}$ are standardized in length. When $\mathbf{A}(b,t) = f^W(b,t)$ then vector element may have widely differing lengths depending on $p(b)$ and $p(t)$.

One advantage of the cosine is that it ranges between -1 and 1, and so removes any arbitrary scaling induced by the range of A and the number of elements in B. When A is simple co-occurrence the cosine is also less sensitive than Euclidean distance to extreme values induced by widely differing basis element frequencies, although a good choice of A should avoid this problem.

The interpretation of similarity as a 'second order' regularity can motivate yet another plausible similarity measure. We may take the correlation coef£cient (Pearson's r) as a measure of how well the elements of each word's vector match. The only difference between this and the cosine measure is that the mean of each vector is included in the similarity measure. This will not only offset the effect of different vector element magnitudes, but also place all calculations in a regular statistical framework. The statistical implications of taking correlation coef£cients over log-odds-ratios remain to be worked out. In addition, all the measures described here will bene£t from a characterization of their properties in small samples. This is future work.

M : Model

A semantic space is fully functional when a B, A and S have been speci£ed. However, it is possible to build a more structured mathematical or statistical model. In LSA the model consists of a projecting vectors into a linear subspace of B using singular value decomposition. This is equivalent to selecting the k orthogonal axes that account for most variance of words in semantic space. Each word is then projected into the the subspace, and point is then 're-in¤ated' back into the full dimensionality and cosine measures applied. Cosines can be taken in the linear subspace without subsequent re-in¤ation as suggested by Berry et al. (1995).

The theoretically important point about LSA's dimensionality reduction is that it is a simple instance of inferring latent structure in distributional data. Parts of speech, and grammatical structures are also examples of

latent structure in the sense that they are in-principle unobservable aspects of words that re¤ect their distributional properties. One important direction for semantic space research is to £nd an appropriate type of latent structure to explain the distributional regularities that are assumed to underly semantic similarity. Biologically motivated models using topographic mapping, and strictly random mappings have also been investigated (Lowe, 2000a,b).

Conclusion

In this paper we have put forward some theory for semantic space models. In addition to presenting a framework for thinking about current semantic space models we have examined the implications of various design choices, emphasized the importance of avoiding frequency biases, and presented methods for doing so. We have also connected semantic space theory to lexicographic methods and to standard problems of bias and variance discussed in the statistical literature.

Acknowledgments

Thanks to Daniel Dennett at the Center for Cognitive Studies at Tufts, where much of the work reported here was done, to the Center for Basic Research in the Social Sciences at Harvard for support, and to three anonymous referees for helpful comments.

References

Agresti, A. (1990). *Categorical Data Analysis*. John Wiley and Sons.

Berry, M. W., Dumais, S. T., and O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4).

Bishop, Y. M. M., Feinberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press.

Burgess, C., Livesay, K., and Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, (25).

Cann, R. (1996). Categories, labels and types: Functional versus lexical. Edinburgh Occasional Papers in Linguistics EOPL-96–3, University of Edinburgh.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, (16).

Dunning, T. (1993). Accurate methods for the statistics for surprise and coincidence. *Computational Linguistics*, (19).

Finch, S. (1993). *Finding Structure in Language*. PhD thesis, Centre for Cognitive Science, University of Edinburgh.

Firth, J. R. (1968). A synopsis of linguistic theory. In Palmer, F. R., editor, *Selected Papers of J. R. Firth: 1952-1959*. Longman.

Foltz, P. W., Kintsch, W., and Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, (25).

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1).

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of induction and representation of knowledge. *Psychological Review*, (104).

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, (25).

Levy, J. and Bullinaria, J. (2000). Learning lexical properties from word usage patterns. In *Proceedings of the 7th Neural Computation and Psychology Workshop*. Springer Verlag.

Lowe, W. (2000a). *Topographic Maps of Semantic Space*. PhD thesis, Institute of Adaptive and Neural Computation, University of Edinburgh.

Lowe, W. (2000b). What is the dimensionality of human semantic space? In *Proceedings of the 7th Neural Computation and Psychology Workshop*. Springer Verlag.

Lowe, W. and McDonald, S. (2000). The direct route: Mediated priming in semantic space. In Gernsbacher, M. A. and Derry, S. D., editors, *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, New Jersey. Lawrence Erlbaum Associates.

Lund, K., Burgess, C., and Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mandelbrot, B. (1954). Structure formelle des textes et communication. *Word*, (10).

McDonald, S. and Lowe, W. (1998). Modelling functional priming and the associative boost. In Gernsbacher, M. A. and Derry, S. D., editors, *Proceedings of the 20th Annual Meeting of the Cognitive Science Society*, New Jersey. Lawrence Erlbaum Associates.

McKoon, G. and Ratcliff, R. (1992). Spreading activation versus compound cue accounts of priming: Mediated priming revisited. *Journal of Experimental Psychology: Learning, Memory and Cognition*, (18).

Rehder, B., Schreiner, M. E., Wolfe, M. B., Laham, D., Landauer, T. K., and Kintsch, W. (1997). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes*, (25).

Zipf, G. K. (1949). *Human Behavior and the Principal of Least Effort*. Addison Wesley.

Individual Differences in Reasoning about Broken Devices: An Eye Tracking Study

Shulan Lu (shulanlu@hotmail.com)
Department of Psychology, University of Memphis, 3693 Norriswood
Memphis, TN 38152 USA

Brent Olde (baolde@memphis.edu)
Department of Psychology, University of Memphis, 3693 Norriswood
Memphis, TN 38152 USA

Elisa Cooper (culdihavben@hotmail.com)
Department of Psychology, University of Memphis, 3693 Norriswood
Memphis, TN 38152 USA

Arthur C. Graesser (a-graesser@memphis.edu)
Department of Psychology, University of Memphis, 3693 Norriswood
Memphis, TN 38152 USA

Abstract

We examined the questions college students ask when everyday devices malfunction. Our investigation of the explanatory reasoning processes is organized around the central theme of question asking. PREG, a model of human question asking, predicts when and what types of questions are asked by humans while comprehending expository texts. PREG has two salient predictions. First, deep comprehenders should ask questions that converge on plausible faults. Second, eye movements should converge on those likely faults. An eye tracking study supported these predictions. The present research supports the claim that question asking and eye tracking are two excellent indicators of device comprehension in the context of breakdown scenarios.

Introduction

Deep comprehension of everyday devices can be manifested in a number of tasks and measures. For example, most colleagues would agree that the deep comprehenders construct coherent representations of functioning devices, draw appropriate inferences, answers explanation-based questions correctly (e.g., why, how, what-if, what-if-not), and solve transfer problems that apply their understanding (Gentner & Stevens, 1983; Graesser, Singer, & Trabasso, 1994). This study investigates the manifestations of deep comprehension of devices that are not as widely recognized. We believe that deep comprehension is required when devices breakdown, when eyes fixate on likely faults, and when questions are asked about likely faults. (Graesser, Olde, & Lu, in press; Graesser, Olde, Pomeroy, Whitten, Lu, & Craig, in press).

Question asking and its role in understanding texts and stories is well-documented (Graesser & McMahen, 1993; Kass, 1992; Schank, 1986, 1999; Ram, 1994). The literature has consistently suggested that the understanding of a story is achieved by identifying the questions raised by the story and then searching for the answers in the story. Question asking has many potential functions in reasoning and problem solving. For instance, question asking is often affiliated with searches, comparisons, explanations, predictions and several other cognitive processes. It is reasonable to say that comprehenders would have more questions as they reason through an expository text and that their questions would manifest comprehension depth. For example, it is very hard to imagine that a shallow comprehender could ask questions addressing the critical causal components of an event. In this article, we examine a model of question asking in the context of understanding technical texts. More specifically, we are interested in the model's predictions about deep versus shallow comprehension of everyday devices as revealed by question asking. It is suggested that question asking opens a window for viewing the sub-processes involved in understanding (Ram, 1994), such as the retrieval of explanations in long term memory and the search of information from a display. However, there has been very little empirical research that documents the relationships among comprehension, question asking, and information search (as reflected in eye tracking). Therefore, we conducted a study in which eye tracking was measured while college students generated questions when confronted with a breakdown scenario.

PREG: A Model of Question Asking

Several computational models of question asking have been constructed in the context of story understanding. These models were capable of generating questions with respect to the goals and sub-goals of a story. However, an adequate model of question asking should be capable of predicting when and what types of questions humans ask as they comprehend expository texts as well as other types of learning material. Otero and Graesser (in press) developed a PREG model of question asking that attempts to capture these question asking mechanisms in detail. The general assumption of the model is that clashes between text input and a reader's existing world knowledge trigger question generation (Graesser, Olde, Pomeroy, et al., in press; Otero et al., in press). Questions are constructed when readers come across information in a text that presents contradictions, anomalies, obstacles to goals, discrepancies, constrasts and other triggers of potential cognitive disequilibrium (Graesser et al., 1993; Graesser & Person, 1994).

The discrepancies between input and world knowledge can be associated with the different levels of representations, ranging from shallow to deep (Britton & Graesser, 1996; Gentner et al., 1983; Graesser, Millis, & Zwaan, 1997; Kieras & Bovair, 1984; Kintsch, 1998). The surface code, which is at the shallowest level, keeps the wording and syntax of a text in a verbatim form. As for the visual modality, it preserves the low-level lines, angles, sizes, shapes, and textures of the picture. The textbase, which is at the intermediate level, in essence is a propositional representation that maintains the meaning of the explicit text and the pictures. The mental model, which is at the deepest level, captures the referential content of the text. When applied to everyday devices, this would include:

1. the components of the electronic or mechanical system;
2. the spatial arrangement of components;
3. the causal chain of events when the system successfully unfolds;
4. the mechanisms that explain each causal step;
5. the functions of the device and device components;
6. the plans of agents who manipulate the system for various purposes.

Quite clearly, a rich set of knowledge structures needs to be constructed when an adult comprehends a device at a deep level.

According to the PREG model, conceptual graph structures are adopted to encode a chronology of events and states that happen during the course of device motion. The conceptual graph structure is not built arbitrarily, but is comprised of a set of categorized nodes which denote concepts and proposition-like descriptions in the text and corresponding visual-spatial information. These nodes are connected by arc categories such as ENABLE, CAUSE, PROPERTY, REASON and OUTCOME. In addition, most arcs are directed with a source node and an end node.

It is assumed that pictorial and textual information is incorporated in a single underlying representation. Empirical studies show that most readers are capable of alternating between picture and text and that the text dominates the reading process when illustrated texts are comprehended (Bagget & Graesser, 1995; Hegarty & Just 1993).

Individual Differences in Question Asking

The current study examines the questions that college students ask when an everyday device malfunctions. For example, consider a cylinder lock and the following breakdown scenario: the key turns, but the bolt does not move. According to PREG, understanding is manifested when a device breaks, not when it is running smoothly. Thus PREG predicts that deep comprehenders should ask good questions that converge on likely faults. More specifically, these questions should tap the nodes in the conceptual graph structure that are the plausible causes of the malfunction. To test this hypothesis, Graesser, Olde, Pomeroy et al. (in press) conducted a study in which 108 participants first read an illustrated text, then were provided a breakdown scenario, and then generated questions. After completing the question asking task, an objective comprehension test on the devices were administered. A battery of tests that measure cognitive abilities and personality were administered in the end.

The results confirmed the hypothesis: Good comprehenders generate high quality questions that focus on plausible faults of the breakdown. Follow-up multiple regression analysis further suggested that ASVAB (the Armed Services Vocational Aptitude Battery, Department of Defense, 1983) technical score was the primary predictor of both deep comprehension and question quality.

Given that technical scientific knowledge turned out to be a robust predictor of device comprehension, we conducted a qualitative analysis of the questions asked by participants with high (upper 33% of distribution) versus low technical knowledge (lower 33% of the distribution). The questions generated by participants with high technical scores tend to converge on the fault components and address the causal connections between parts, processes, and relations that are in the chain of breakdown. The questions asked by low technical participants tend to be diffuse. That is, most of the components in a system were addressed in the hope that it might turn out to be pertinent instead of converging on 1 or 2 parts. Their questions rarely were

elaborations on the causal links addressing the malfunction.

Since the above patterns emerged, we were curious to know whether there were systematic differences in the eye movement patterns between individuals with different levels of cognitive abilities. That is the focus of the present study.

Question Asking and Control of the Eye

Eye movements provide an important window for understanding the cognitive processes and representations that play a role in a particular cognitive task. However, no one has investigated the relationship between eye movements and the cognitive components in question asking. PREG predicts that eye movements should converge on the likely faults. As far as individual differences in the eye movement are concerned, the following hypotheses could be directly generated from the PREG model. First, deep comprehenders are expected to have a high density of eye fixations occur at words, objects, parts, and processes that are at the source of cognitive disequilibrium (e.g., anomalies, contradictions, broken parts, contrasts, missing components, and so on), while shallow comprehenders should indiscriminately scan the regions of the illustrated text. A sufficient amount of technical knowledge is necessary for identifying anomalies in a system. Thus, technical knowledge and other indices of deep comprehension should be positively correlated with measures of the fixations that assess the extent to which a comprehender focuses on fault areas.

Method

Participants

The participants were forty college students at the University of Memphis. The students participated for course credit in an introductory psychology class.

Illustrated Texts and Question Asking Tasks

The participants read 5 illustrated texts on everyday devices: a cylinder lock, an electronic bell, a car temperature gauge, a toaster, and a dishwasher. The illustrated texts were extracted from Macaulay's (1988) book, *The Way Things Work.* These were the same devices that were used in the Graesser, Olde, Pomeroy et al., (in press) study, except that the clutch was dropped from the current study; it was extremely difficult for participants to differentiate and label the individual teeth in the wheels of the clutch mechanism.

As in the Graesser, Olde, Pomeroy et al., (in press) study, there were five trials, each of which consisted of two phases. The participant first read an illustrated text for 3 minutes, which was displayed on a computer monitor. After the reading phase, the breakdown description was presented either above or to the left of the illustrated text and the participants began the question asking phase (while the illustrated text remained on the screen). The participants asked questions aloud for 90 seconds during this phase and the protocol was recorded. The previous study had participants generate questions in writing whereas the present study collected spoken questions. Each participant furnished question asking protocols for all 5 devices. The assignment of devices to test order was counterbalanced across the 40 participants with a Latin square.

Device Comprehension Test

The participants subsequently completed a 30-question test of device comprehension (5 devices x 6 three-alternative, forced-choice question per device). All 30 questions were generated from a theoretical framework in qualitative physics (Forbus, 1984). Suppose there are N components in a system and their states are delineated as either inhibitory, excitatory, or neutral as affected by other components in the causal network. The test questions are concerned with how tweaking one component A has an impact on another component B. An example question constructed according to the three possible states is as follows:

What happens to the pins when the key is turned to unlock the door?

(a) they rise
(b) they drop
(c) they remain stationary (correct answer)

It is not likely that participants will be able to answer such questions correctly without deep comprehension of the devices. It is reasonable to predict that deep comprehenders will be able to trace the causal antecedents and causal consequences of the events (Graesser & Bertus, 1998) and shallow comprehenders will lose track of causal connections. The device comprehension test was thus designated as the gold standard for deep comprehension. It is predicted that performance on the device comprehension test should positively correlate with the quality of the questions that get asked and also with fixations on the faults of a breakdown scenario. The device comprehension score could vary from 0 to 30. A score of 10 would be chance performance if there were no sophisticated guessing or background knowledge.

Battery of tests of individual differences

The participants completed assessments of the following measures of individual differences: four scales of technical knowledge (mechanical comprehension, electronics, general science, auto & shop) extracted from ASVAB (Department of Defense, 1983), and additional tests of spatial reasoning (Bennet,

Seashore, & Wesman, 1972) and openness (Costa & McCrae, 1991). These were included, as they were the statistically significant predictors of deep comprehension and question asking in the Graesser, Olde, and Pomeroy et al. (in press) study.

Recording of eye tracking and question asking

Eye movements were recorded by a Model 501 Applied Science Laboratory eye tracker. There was a magnetic head tracker so the participants could move the head during data collection. The participants were calibrated before they started the experimental session of reading the illustrated texts and asking questions. During calibration, the participants viewed 9 points on the computer display and a computer recorded the x-y coordinates. The calibration process took 10-15 minutes. Participants were dismissed if they wore glasses, but the equipment could accommodate contact lenses.

The experimental session was videotaped and audio recorded. The VCR camera focused on a scene monitor screen, on which the illustrated text being viewed by the participant and the trace of the participant's eye movements were mirrored. The VCR recorded both the illustrated text and a superimposed image of what the left eye was focusing on. The superimposed image showed the locus of (a) the focus of the eye and (b) an X-Y axis with the 0-0 point at the center of the focus. The voice of the participant was recorded on the VCR so that the spoken questions could be transcribed. This set-up allowed us to record and review (a) the contents of the computer display, (b) the focus of the left eye, and (c) the voice of the student asking questions.

Computer software was available to record eye movements at a fine-grained level. The software produces area plots for specific areas of interest. In particular, we were interested in the portions of eye fixations focusing on the areas of interest associated with faults. These faults were sometimes in the text and sometimes in the picture.

The following measures were scored on the think aloud protocols collected in the question asking task.

Volume of questions: The number of questions that were asked in the question asking task.
Question Quality: The number or proportion of questions that referred to a plausible explanation of the breakdown.

Trained judges coded the verbal protocols with an acceptable level of reliability.

Results and Discussion

Descriptive Statistics

Table 1 presents means and standard deviations for the measures collected in this study. The ASVAB measures were comparable to normal college student populations. Scores of spatial reasoning were not significantly different from the scores for college students reported in Bennet et al. study (1972).

Table 1: Descriptive Statistics on Measures of Individual Differences

Measures	Mean	SD
Mechanical	13.9	6.0
Electronics	9.3	4.3
Auto & Shop	10.0	5.3
General Science	16.9	4.7
Spatial	23.8	15.2
Openness	52.1	9.2
Gender	1.25	.44

The Coordination between Question Asking and Eye Movements

It is important to examine the device comprehension scores first. The mean device comprehension score was 18.6 (SD = 4.6), which is 62% of the questions being answered correctly. As would be expected, device comprehension was significantly correlated with the number of fault questions asked ($r = .45$, $p < .01$). In addition, device comprehension showed a high positive correlation with all the ASVAB measures, in particular, ASVAB technical knowledge ($r = .54$) and general science ($r = .60$).

The data supported PREG's predictions. There were 29.5 fixations (SD = 11.1) on plausible faults per device, or 9.3 seconds (SD = 3.9) out of 90 seconds. 11.5% of the eye fixations and 10.4% of the time focused on the fault areas. The index of deep comprehension, i.e., device comprehension score, was significantly correlated with the eye tracking measure of number of fixations on faults ($r = .43$).

We examined the role of question asking in eye tracking at this point. The data showed strong evidence that there was consistent coordination between question asking and eye movements. Furthermore, there were significant differences in the measures of eye tracking between participants who had a relatively high number of fault questions and those who had a low number of questions, e.g., number of fixations on faults (167.6 with SD = 42.2 versus 127.5 with SD = 61.0, for high versus low).

Given the importance of cognitive abilities during device comprehension, we explored which measures of cognitive ability are more capable of discriminating deep comprehension versus shallow comprehension as measured by question asking and eye movements. General science turned out to be a significant predictor of the eye tracking measure and the number of fault questions asked. The participants who scored higher on general science had more questions on faults (7.7 with

SD = 2.7 versus 6.0 with SD = 3.0, for high versus low), and had bigger number of fault fixations (173.1 with SD = 48.7 versus 122.0 with SD = 50.9).

Openness is one of the "big five" personality factors: neuroticism, extroversion, openness, agreeableness, and conscientiousness (Costa et al., 1991). The subscale of openness attempts to capture creativity. A t-test on the eye tracking measure and question asking measure showed that there were significant differences between participants with high openness scores versus those with low openness scores: number of fixations on faults (167.8 with SD = 49.6 versus 127.3 with SD = 55.0, for high versus low), and number of fault questions (8.2 with SD = 2.9 versus 5.5 with SD = 2.4). The data suggested that openness was a robust predictor of question asking and subsequently affected the patterns of eye movements.

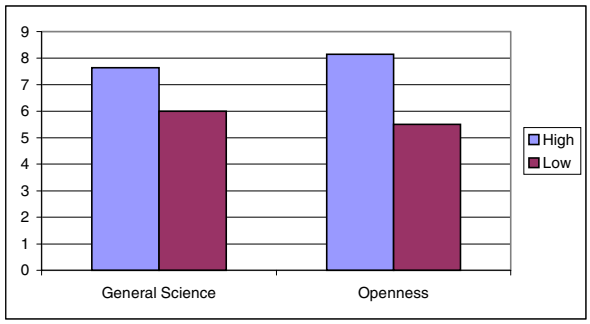


Figure 1: The number of questions on faults by high versus low on general science and openness scales.

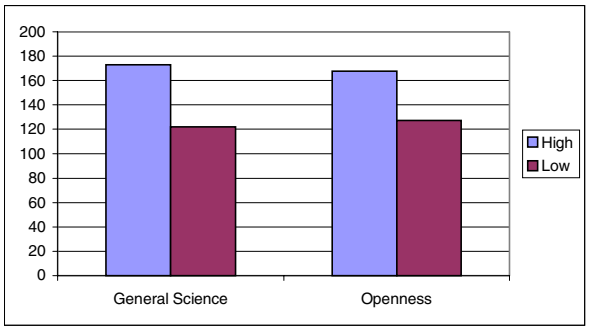


Figure 2: The number of fixations addressing faults by high versus low on general science and openness scales.

There have been two different views concerning the relationships between eye movements and ongoing cognitive processes (Rayner, Reichle, & Pollatsek, 1998). One is that eye movements are mainly driven by oculomotor processes; the other is that there is a correspondence between eye movements and cognitive processes. The data quite clearly suggested the coordination between question asking and eye movements. The question at this point was to what extent question asking was in control of eye movements. The ideal method of addressing this question would be structural equation modeling. However, given the number of participants in the present study, we could only conduct a partial correlation analysis.

If our PREG model is correct, there should not be a significant correlation between the number of questions addressing faults and readers' technical knowledge after partitioning out the variance of question asking. Subsequent data analysis supported the hypothesis that question asking is in control of eye movements rather than eye tracking guiding question asking. When the variable (number of fault questions) was controlled, the correlation between technical knowledge and the number of fixations on faults approached 0 ($r = .14$, $p < .05$). However, controlling the variance of eye movements did not affect the correlation between technical knowledge and the number of fault questions. They remained significantly correlated ($r = .39$, $p < .05$). The results are consistent with some recent findings which suggest that some cognitive processes are fast enough to affect eye movements (Rayner et al., 1998).

In short, it appears that individuals with high scores on general science and openness are most capable of asking questions about the anomalous information in a system. Subsequently these individuals move their eyes to the plausible fault areas and verify their reasoning about the breakdown scenario. On the other hand, individuals, who are low on the general science and openness scales, tend to be less sensitive to the contradictions when they arise. Thus they resort to the strategy of scanning all the regions of a text in the hope of hitting the target.

Conclusion

The current research has demonstrated the usefulness of the eye tracking data for studying the cognitive components in question asking. The analysis of eye fixations provides an account of how people with different levels of cognitive ability and different types of personality reason through the device malfunction and how their explanatory reasoning processes center around anomaly detection and question asking.

Acknowledgements

This research was funded by the Office of Naval Research (N00014-98-1-0331). We thank Victoria Pomeroy, and Shannon Whitten for collecting and analyzing data on this project.

References

Baggett, W. B., & Graesser, A. C. (1995). Question answering in the context of illustrated expository text. *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 334-339). Hillsdale, NJ: Lawrence Erlbaum.

Bennet, G. K., Seashore, H. G., & Wesman, A. G. (1972). *Differential aptitude test: Spatial relations., Form T*. New York: Psychological Corporation.

Britton, B., & Graesser, A. C. (1996) (Eds.). *Models of understanding text*. Hillsdale, NJ: Erlbaum.

Costa, P. T., & McCrae, R. R. (1991). *NEO: Five Factor Inventory*. Odessa, FL: Psychological Assessment Resources.

Department of Defense (1983). *Armed Services Vocational Aptitude Battery, Form 12a*. Washington, D.C.: Department of Defense.

Forbus, K. (1984). Qualitative process theory. *Artificial intelligence, 24*, 85-168.

Gentner, D., & Stevens, A. L. (1983) (Eds.). *Mental models*. Hillsdale, NJ: Erlbaum.

Graesser, A. C., Bertus, E. L. (1998). The construction of causal inferences while reading expository texts on science and technology. *Scientific Studies of Reading, 2*, 247-269.

Graesser, A. C., & McMahen, C. L. (1993). Anomalous information triggers questions when adults solve quantitative problems and comprehend stories. *Journal of Educational Psychology, 85*, 136-151.

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal, 31*, 104-137.

Graesser, A. C., Singer, M., Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371-395.

Graesser, A. C., Millis, K. K., Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology, 48*, 163-189.

Graesser, A. C., Olde, B., & Lu, S. (in press). Question-driven explanatory reasoning about devices that malfunction. In T. Filjak (Ed.), *Proceedings of the 36th International Applied Military Psychology Symposium*.

Graesser, A. C., Olde, B., Pomeroy, V., Whitten, S., Lu, S., & Craig, S. (in press). Inferences and questions in science text comprehension. In J. Otero and M. Helena (Eds.), *Science text comprehension*.

Hegarty, M., & Just, M. A. (1993). Constructing mental models of machines from text and diagrams. *Journal of Memory and Language, 32*, 717-742.

Kass, A. (1992) Question-asking, artificial intelligence, and human creativity. In T. Lauer, E. Peacock, & A. C. Graesser, (Eds.), *Questions and information systems.* Hillside, NJ: Erlbaum.

Kieras, D. E., & Bovair, S. (1984). The role of a mental model in learning to operate a device. *Cognitive Science, 8*, 255-274.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York, NY: Cambridge University Press.

Macaulay, D. (1988). *The way things work*. Boston, MA: Houghton Mifflin.

Otero, J., & Graesser, A. C. (in press). PREG: Elements of a model of question asking. *Cognition and Instruction*.

Ram, A. (1994) AQUA: Questions that drive the explanation process. In R.C. Shank, A. Kass, & C.K. Riesbeck (Eds.), *Inside Case-Based Explanation*. Hillsdale, NJ: Erlbaum.

Rayner, K, Reichle, E. D. & Pollatsek, A. (1998). Eye movement control in reading: An overview and model. In G. Underwood (Eds.), *Eye guidance in reading and scene perception*. Oxford, England: Elsevier.

Schank, R. C. (1986). *Explanation patterns: Understanding mechanically and creatively*. Hillsdale, NJ: Erlbaum.

Schank, R. C. (1999). *Dynamic memory revisited*. Cambridge: Cambridge University Press.

Modeling Forms of Surprise in an Artificial Agent

Luís Macedo (lmacedo@isec.pt)
Instituto Superior de Engenharia de Coimbra / CISUC – Centro de Informática e Sistemas da Universidade de Coimbra, Quinta da Nora
3031-601 Coimbra, Portugal

Amílcar Cardoso (amilcar@dei.uc.pt)
Departamento de Engenharia Informática da Universidade de Coimbra / CISUC – Centro de Informática e Sistemas da Universidade de Coimbra, Pinhal de Marrocos
3030 Coimbra, Portugal

Abstract

Mainly rooted in the cognitive-psychoevolutionary model of surprise proposed by the research group of the University of Bielefeld (Meyer, Reisenzein, Schützwohl, etc.), the computational model of surprise described in this paper relies on the assumption that surprise-eliciting events initiate a series of mental processes that begin with the appraisal of unexpectedness, continue with the interruption of ongoing activity and the focusing of attention on the unexpected event, and end with the analysis and evaluation of that event plus revision of beliefs. With respect to the computation of unexpectedness, the model also incorporates suggestions by Ortony and Partridge. This model of surprise is implemented in an artificial agent called S-EUNE, whose task is to explore uncertain and unknown environments. The accuracy of our surprise model was evaluated in a series of experimental tests that focused on the comparison of surprise intensity values generated by the artificial agent with ratings by humans under similar circumstances.

Introduction

Roughly speaking, artificial and biological agents accept percepts from the environment and generate actions. Since different actions may lead to different states of the world, in order to perform well (to execute the "right" action), some kinds of artificial agents make use of a mathematical function that maps a state of the world onto a real number - the utility value. Thus, in those agents, decision-making is performed by selecting the action that leads to the state of the world with the highest utility (Russell & Norvig, 1995; Shafer & Pearl, 1990).

Although research in Artificial Intelligence has all but ignored the significant role of emotions in reasoning/decision-making (e.g., Damásio, 1994), several computational models for emotions have been proposed in the past years, based in part on research in psychology and neuroscience (for a detailed review of those models see e.g., Pfeifer, 1988; Picard, 1997).

Considered by many authors as a biologically fundamental emotion (e.g., Ekman, 1992; Izard, 1977), *surprise* may play an important role in cognitive activities, especially in attention focusing and learning (e.g., Izard, 1977; Meyer, Reisenzein, & Schützwohl, 1997; Ortony & Partridge, 1987; Reisenzein, 2000b) (note however, that some authors, like Ortony, Clore, and Collins, 1988, do not consider surprise an emotion). According to the research group of the University of Bielefeld, Germany (e.g., Meyer et al., 1997), surprise has two main functions, informational and motivational, that together promote both immediate adaptive actions to the surprising event and the prediction, control and effective dealings with future occurrences of the event. Ortony and Partridge's view of surprise shares aspects with this model, especially in that both assume that surprise is elicited by unexpected events. The same is also true for Peters' (1998) computational model of surprise, implemented in a computer vision system, that focuses on the detection of unexpected movements.

In this paper, we propose a computational model for surprise that is an adaptation (although with several simplifications) of the models proposed by the German research group of the University of Bielefeld and by Ortony and Partridge.

The following section presents an overview of the overall agent's architecture into which the surprise model is integrated. Subsequently, we explain this model in detail. Finally, we describe experimental tests carried out to evaluate the accuracy of the surprise model.

Overview of the Agent's Architecture

EUNE (Emotional-based Exploration of UNcertain and UNknown Environments) is an artificial agent whose goal is the exploration of uncertain and unknown environments comprising a variety of objects, and whose behavior is controlled by emotions, drives and other motivations. Besides desiring to know or be aware of the objects belonging to the environment, EUNE is also able to "feel" the emotions (including surprise)

those objects cause. In fact, these "felt emotions" guide the exploratory behavior of EUNE: roughly speaking, at any given time, among several objects available in the environment, EUNE selects that object for study and analysis that causes more positive emotion and less negative emotion (Izard, 1977) (see Reisenzein, 1996, for related theories of emotional action generation, and Barnes & Thagard, 1996, for an alternative approach to emotional decision-making). This process is repeated until all objects in the environment have become known.

In this article, we describe S-EUNE, a simplified version of EUNE whose emotional makeup is confined to the emotion of surprise. As many other agents, S-EUNE has percepts, actions, goals, memory, emotions/drives, and deliberative reasoning/decision-making (Figure 1) (for more details on this architecture see Macedo & Cardoso, 2001).

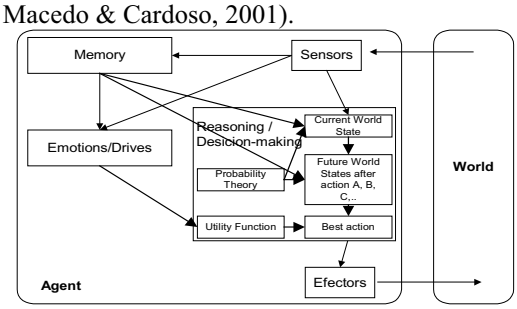


Figure 1: S-EUNE's architecture.

Previously defined by the user, the environment comprises a variety of objects located at specific positions. In the present article, these objects are confined to buildings. Each object comprises three distinct, fundamental components: *structure*, *function* and *behavior* (Goel, 1992). For the sake of simplicity, the *structure* (the visible part of the object), is restricted to the shape of the object (e.g., triangular, rectangular, etc.); however, any object may comprise several sub-objects. The *function* of the object concerns its role in the environment (e.g., house, church, hotel, etc.). The *behavior* of the object concerns its activity (actions and reactions) in response to particular features of external or internal stimuli (e.g., static, mobile).

The perceptual system of the agent (two simulated sensors) provides information related to the *structure*, the *function*, and the *behavior* of the objects, as well as the distance of the objects. Note that the *function* of the objects is not accessible (i.e., cannot be inferred from visual information) unless the agent is at the same place as the object.

As a knowledge-based agent, S-EUNE stores all the information acquired through the sensors in its memory unit. The agent's knowledge base is of an episodic kind: each object is stored, in the form of a graph, as a separate case in episodic memory. In addition, each object representation is associated with a number that expresses its absolute frequency (Figure 2).

Field \ Case	C_1	C_2	C_3	C_4
Structure				
Function	House	House	Church	Hotel
Behavior	Static	Static	Static	Static
Abs. Freq.	50	40	5	5

Figure 2: Example of the episodic memory of S-EUNE after exploring an environment.

When information from the environment is sampled, the *surprise generation module* compares that information to the information stored in memory and outputs the intensity of the elicited surprise. A corresponding facial expression is also produced. The model of surprise is described in more detail in a later section. Note that, in the more general EUNE agent, this component - in this case called the *emotion, drives and other motivations module* - may also comprise other emotions apart from surprise, such as anger, sadness, etc., as well as drives and other motivations.

The *reasoning/decision-making module* of S-EUNE receives the information from the simulated external world and outputs the action that has been selected for execution. This module comprises several subprocesses: (i) taking the information of the world provided by the sensors (which may be incomplete) as input, the current state of the world (the agent's current position, the position of the objects, etc.) is computed; (ii) taking the current state of the world, probability theory and memory-stored information as input, possible future world states and respective probabilities are computed for the actions that the agent can perform; (iii) from these actions, a single one (presumably the best one) is selected. This is the action that maximizes the Utility Function (Russell & Norvig, 1995; Shafer & Pearl, 1990), which in the case of S-EUNE relies heavily on the anticipated intensity of surprise elicited by the future state of the world. Thus, the preferences of S-EUNE are reflections of its anticipated surprise. In order to achieve this goal of maximizing anticipated surprise, the *reasoning/decision-making module* makes use of an Utility Function, abbreviated *U(W)*, which is based on the surprise function (defined in the next section) as follows:

$$U(W) = f(U_{surprise}(W)) = f(SURPRISE(Agt, Obj(W)))$$

This Utility Function means that the utility of a world state W is given by the surprise that the state W causes the agent to "feel". In this article, a world state is defined as "being close to or seeing an object" (the object that is currently the focus of attention of the agent's sensors), and f is taken to be the identity function, implying that *U(W)* increases monotonically

with the intensity of surprise. As a consequence of this, the agent always selects for approach the object that actually elicits and/or promises to elicit maximum surprise.

Surprise Model

As mentioned before, our model of surprise is mainly based on Ortony and Partridge's proposals and on the University of Bielefeld model. We will now give an overview of these models and then explain our computational model by comparing it with these two models.

Background Models

Ortony and Partridge (1987) proposed that there is a difference between surprisingness and expectation failure. They suggest that, although surprise sometimes results from expectation failure, it is frequently also caused by events for which expectations were never computed. In other words, one can be surprised by something one didn't expect without having to expect something else. Ortony and Partridge also proposed that surprisingness is an important variable in artificial intelligence systems, particularly for attention and learning.

The following assumptions were made in their model: the system (or agent) receives an input proposition; the system has an episodic and semantic memory; elements of the memory may be immutable (propositions that are believed to be always true) or typical (those that are believed to be sometimes true); and, some elements of the memory are activated when an input proposition is received.

Ortony and Partridge further distinguish between *practically deducible propositions* and *practically non-deducible propositions*. *Practically deducible propositions* comprises the propositions that are explicitly represented in memory, as well as those that can be inferred from them by few and simple deductions. Hence, *practically deducible propositions* are that subset of formally deducible propositions that don't require many and complex inferences. Furthermore, *practically deducible propositions* may be actively or passively deduced in a particular context. In the former case, their content corresponds to *actively expected* or *predicted* events; in the latter case, to *passively expected* (*assumed*) events.

Based on these assumptions, Ortony and Partridge proposed that surprise may result from three situations (Table 1 presents the correspondent range of values): (i) *active expectation failure*: here, surprise results from a conflict or inconsistency between the input proposition and an *active prediction* or *expectation*; (ii) *passive expectation failure* (or *assumption failure*): here, surprise results from a conflict or inconsistency between the input proposition and what the agent implicitly knows or believes (*passive expectations* or *assumptions*); and (iii) *unanticipated incongruities* or deviations from norms: here, surprise results from a conflict or inconsistency between the input proposition (which in this case is a *practically non-deducible proposition*) and what, after the fact, may be judged to be normal or usual (cf. Kahneman & Miller, 1986), that is, *practically deducible propositions* (immutable or typical) that are suggested by the unexpected fact. Note that, in this case, at least prior to the unexpected event, there are no expectations (passive or active) with which the input proposition could conflict.

Table 1: Three different sources of surprise and correspondent range of values (adapted from Ortony & Partridge, 1987).

Confronted proposition	Related Cognition	
	Active	Passive
Immutable	[1]; $S_A=1$; *Prediction*	[2]; $S_P=1$; *Assumption*
Typical	[3]; $0< S_A<1$; *Prediction*	[4]; $S_P<S_A$; *Assumption*
Immutable	[5]; ∅	[6]; $S_P=1$; *none*
Typical	[7]; ∅	[8]; $0< S_P<1$; *none*

In their cognitive-psychoevolutionary model, the research group of the University of Bielefeld has made similar assumptions as Ortony and Partridge, namely that surprise (considered by them as an emotion) is elicited by the appraisal of unexpectedness. More precisely, it is proposed that surprise-eliciting events give rise to the following series of mental processes: (i) the appraisal of a cognized event as exceeding some threshold value of unexpectedness (schema-discrepancy) - according to Reisenzein (1999), this is achieved by a specialized comparator mechanism, the unexpectedness function, that computes the degree of discrepancy between "new" and "old" beliefs or schemas; (ii) interruption of ongoing information processing and reallocation of processing resources to the investigation of the unexpected event; (iii) analysis/evaluation of that event; (iv) possibly, immediate reactions to that event and/or updating or revision of the "old" schemas or beliefs.

Our Computational Model of Surprise

We have implemented a computational model of surprise, in the context of S-EUNE, that is an adaptation (although with some simplifications) of the University of Bielefeld's model and in which the above-mentioned four mental processes elicited by surprising events are present. The suggestions by Ortony and Partridge are mainly concerned with the first of these steps, and are compatible with the Bielefeld model (see Reisenzein, 1999). Accordingly, we drew on these assumptions for the implementation of the appraisal of unexpectedness and the computation of the intensity of surprise, as well as the selection of knowledge structures in our model.

Within our model, knowledge is of an episodic kind, rather than being both semantic and episodic (although this will be part of our future work) as in Ortony and Partridge's model. Therefore, the knowledge structure of our model differs also from the schema-theoretic framework of the University of Bielefeld's model, that also assumes both episodic and semantic knowledge. In our model an input proposition (or new belief) is related to a visual object or parts of an object (for instance the visual effect of an object with squared windows, rectangular door, etc.). Besides, the agent has in its episodic memory explicit representations of similar propositions. Following Ortony and Partridge, we also distinguish between *deducible* and *non-deducible*, *active* and *passive*, *immutable* and *typical* propositions as well as between different possible sources of surprise (see Table 1). The immutability of a proposition can be extracted from the absolute frequency values associated with the cases (see Figure 2 above). For instance, the proposition "houses have squared facades" is immutable (since all the houses in memory have squared facades), whereas "houses have squared windows" is a typical proposition with a probability (immutability) value of .55 (as implied by Ortony and Partridge's model, in our model immutability is a continuous variable).

The usual activity of the agent is moving through the environment hoping to find buildings that deserve to be investigated. When one or more buildings are perceived, the agent computes expectations for their *functions* (for instance, "it is a house with 67% of probability", "it is a hotel with 45% of probability", etc.). Note that the *function* of a building is available to the agent only when its position and that of the building are the same. On the basis of this information (the *structure* of the object and predictions for its *function*), the agent then computes the surprise intensity that the building causes through the computation of its degree of unexpectedness (described below). Then, the building with the maximum estimated surprise is selected to be visited and investigated. This corresponds to the "interruption of ongoing activity" assumed in the Bielefeld model of surprise. The previously estimated value of surprise may now be updated with the additional information concerning the *function* of the building. The object is then stored in memory and the absolute frequencies of the affected episodes in memory are updated. This is a simplification of the fourth step of the University of Bielefeld's model (for alternative approaches to belief revision, see, for instance, Gärdenfors, 1988). Note that the experience of surprise is also accompanied by a correspondent facial expression (raised eyebrows, widened eyes, open mouth) (Ekman, 1992).

To see how the first step, the appraisal of unexpectedness, is performed, we now describe how the degree of unexpectedness is computed in the three surprise-eliciting situations distinguished by Ortony and Partridge.

As said above, when the agent sees the structure of a building it computes expectations (*deducible*, *active expectations*) for its *function* (e.g., "it is a hotel with 45% of probability", etc.). If, after visiting that building, the agent finds out that it is a post office, it would be surprised, because its *active expectations* conflict with the input proposition (note that, in our model, belief conflicts may be partial rather as well as total). This is thus an example of the first source of surprise distinguished by Ortony and Partridge. In contrast, when the agent sees a building with a window (or roof, etc.) of a particular shape (for instance, circular), although it may not have made an *active prediction* for its shape, it is able to infer that it expected a rectangular shape with, for instance, 45% probability, a squared shape with 67%, etc. This is an example of a *deducible*, *passive expectation*: although not made before the agent perceived the building, it could easily infer an expectation for the shape of the window after it was perceived. This case is therefore an example of the second source of surprise because the input proposition "has a circular window" conflicts with the agent's *passive expectations*. Finally, when the agent sees a building with no facade, it has neither an *active* nor a *passive expectation* available, because there are no buildings with no facade in its memory and therefore the agent could not predict that. Thus, "the house has no facade" is an example of a *non-deducible proposition*. This is an example of the third source of surprise: there is a conflict between the input proposition "the house has no facade" and what after the fact is judged to be normal or usual ("buildings have a facade").

Let us now describe how the intensity of surprise is computed. There is experimental evidence supporting that the intensity of felt surprise increases monotonically, and is closely correlated with the degree of unexpectedness (see Reisenzein, 2000b, for a review of these experiments). This suggests that unexpectedness is the proximate cognitive cause of the surprise experience. On the basis of this evidence, we propose that the surprise felt by an agent *Agt* elicited by an object Obj_k is proportional to the degree of unexpectedness of Obj_k, considering the set of objects present in the memory of the agent. According to probability theory (e.g., Shafer & Pearl, 1990), the degree of expecting that an event X occurs is given by its probability P(X). Accordingly, the improbability of X, denoted by 1-P(X), defines the degree of not expecting X, and the intensity of surprise can, for simplicity, be equated with unexpectedness:

$$SURPRISE(Agt, Obj_k) =$$
$$= DegreeOfUnexpectedness(Obj_k, Agt(Memory)) = 1 - P(Obj_k)$$

Although other probabilistic methods might be used to compute P(X), in the case of objects comprising several components we propose to compute the probability of the whole object Obj_k as the mean of the conditional probabilities of their n constituent parts, which are individually computed using Bayes's formula (Shafer & Pearl, 1990) (note that each one of those conditional probabilities individually gives the degree of unexpectedness of a specific piece of the object, given as evidence the rest of the object):

$$P(Obj_k) = \frac{\sum_{l=1}^{n} P(Obj_k^l \mid Obj_k^1, Obj_k^2, ..., Obj_k^{l-1}, Obj_k^{l+1}, ..., Obj_k^n)}{n}$$

Experimental Tests

Although our model is consistent with the experimental evidence reported, we performed two new experiments to test the following issues: (i) whether the intensity values generated by the artificial agent match those of humans under similar circumstances; (ii) the role of the amount of previous knowledge on the surprise intensity; (iii) whether the surprise intensity values generated by the artificial agent fall within the range of the surprise intensity values proposed in Ortony and Partridge's model. In both experiments, the participants (S-EUNE and 60 humans with mean age of 20.5 years) were presented with 40 quiz-like items. Experiment 1 was performed in an abstract domain with hedonically neutral events (see Stiensmeier-Pelster, Martini, & Reisenzein, 1995, for a similar experiment with humans). Each "quiz item" consisted of several sequences of symbols. Some of the "quiz items" contained a single sequence in which one symbol was missing. Experiment 2 was performed in the domain of buildings. In this case, each "quiz item" consisted of the presentation of a building, and some items did not include information about its *function* (see Reisenzein, 2000a, for a conceptually similar experiment with humans). In those cases where a symbol of the sequence (Experiment 1) or information about the *function* of the building (Experiment 2) was missing, the participants had to state their expectations for the missing symbol or the missing *function*. Subsequently, the "solution" (the missing information) of the "quiz item" was presented and the participants were asked to rate the intensity of felt surprise about the "solution", as well as for the whole sequence/building. For "quiz items" ending with complete sequences or complete buildings, the participants had to rate the intensity of felt surprise about a specified element of the sequence or a specified piece of the building. Subsequently, they also indicated their *passive expectations* for that element/piece. The "quiz items" used in both experiments were selected on the basis of a previous questionnaire study. They were equally distributed among the three sources of surprise described earlier, as well as among different intensities of surprise ranging from low to high.

Figure 3 presents the results of Experiment 1. It can be seen that the intensity of surprise computed for an element of a sequence by the agent (labeled *S-EUNE-Piece* in Figure 3) is close (average difference = .065, i.e., 6.5%) to the corresponding average intensity given by the human judges (*Humans Average-Piece*). Even better results (average difference = .022) were obtained for the surprise values computed for the whole sequence (*S-EUNE-Whole* and *Humans Average-Whole*). Figure 3 also shows that the standard deviations of the surprise intensities given by the 60 humans (*S.D.-Humans-Piece*, *S.D.-Humans-Whole*) were less than .23 (for an element) and .18 (for the whole sequence).

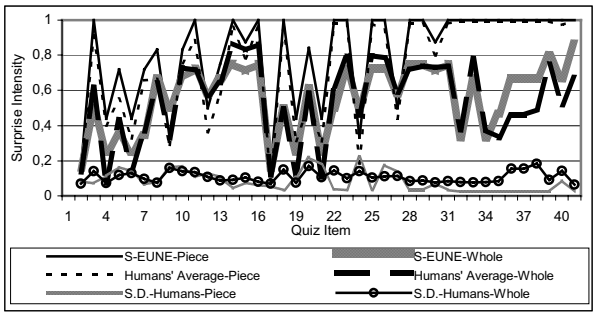


Figure 3: Results of Experiment 1.

Figure 4 presents the results of Experiment 2. In this experiment, S-EUNE answered the "quiz items" several times, each time with a different episodic memory. Due to the lack of space, we reported only the results of three sessions, denoted by S-EUNE-I, IV and V (with I, IV and V denoting an increasingly large memory). It can be seen that the surprise values of the agent are not as close to the human judgments as in the previous domain. For instance, the average differences for S-EUNE-V were .47 (for a piece of a building) and .05 (for the whole building). This happened most likely because, in contrast to the previous, hedonically neutral domain, in the domain of buildings the knowledge of humans and of the agent is different. However, the results suggest that the larger the episodic memory, and the closer its probability distribution corresponds to the real world, the closer are the surprise values given by the agent and by the humans. For instance, S-EUNE-V (*S-EUNE-V-Piece* and *S-EUNE-V-Whole*) showed the best correspondence to the human ratings. This experiment also confirms to some extent the dependence of surprise on the contents and developmental stage of memory, suggested by studies that compared the surprise reactions of adults with those of children (Schützwohl & Reisenzein, 1999).

Both experiments also confirmed that the values of surprise fall in the ranges predicted by Ortony and Partridge, with the exception that, in the case of the source of surprise corresponding to cell [8] of Table 1, the values are always 1, and, in the case of cell [4], $S_P=S_A$.

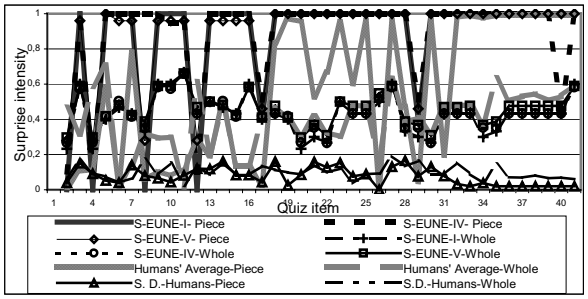


Figure 4: Results of Experiment 2.

Conclusions

The results of the reported experiments suggest that the described computational model is a possible model of surprise. However, alternative surprise functions are conceivable, such as, $SURP(O)=\ln_2(1/P(O))$ (as suggested by information theoretic accounts) or $SURP(O)=1-P(O) \Leftarrow P(O)<.5$; $SURP(O)=0 \Leftarrow P(O)\geq .5$ (as suggested to us by Rainer Reisenzein). We are currently exploring these and other alternatives.

Acknowledgments

We would like to thank Andrew Ortony and Rainer Reisenzein for their helpful comments, and the participants of our experiments for their cooperation.

References

Barnes, A., & Thagard, P. (1996). Emotional decisions. *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 426-429). Mahwah, NJ: Erlbaum.

Damásio, A. (1994). *Descartes'error, Emotion Reason and the Human Brain*. New York: Grosset/Putnam Books.

Ekman, P. (1992). An argument for basic emotions. In N. L. Stein & K. Oatley (Eds.), *Basic Emotions* (pp. 169-200). Hove, UK: Erlbaum.

Gärdenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states*. Cambridge, MA: Bradford Books.

Goel, A. (1992). Representation of design functions in experience-based design. In D. Brown, M. Waldron, & H. Yoshikawa (Eds.), *Intelligent Computer Aided Design* (pp. 283-308). Amsterdam: North-Holland.

Izard, C. (1977). *Human Emotions*. New York: Plenum.

Kahneman, D., & Miller, D. T. (1986). Norm theory: comparing reality to its alternatives. *Psychological Review*, *93*, 136-153.

Macedo, L., & Cardoso, A. (2001). SC-EUNE – Surprise/Curiosity Exploration of UNcertain and UNknown Environments. *Proceedings of the AISB'01 Symposium on Emotion, Cognition and Afective Computing* (pp. 73-81). York, UK: SSAISB.

Meyer, W., Reisenzein, R., & Schützwohl, A. (1997). Towards a process analysis of emotions: The case of surprise. *Motivation and Emotion, 21*, 251-274.

Ortony, A., & Partridge, D. (1987). Surprisingness and Expectation Failure: What's the Difference?. *Proceedings of the 10th International Joint Conference on Artificial Intelligence* (pp. 106-108). Los Altos, CA: Morgan Kaufmann.

Ortony, A., Clore, G., & Collins, A. (1988). *The cognitive structure of emotions*. New York: Cambridge University Press.

Peters, M. (1998). Towards artificial forms of intelligence, creativity, and surprise. *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 836-841). Mahwah, NJ: Erlbaum.

Pfeifer, R. (1988). Artificial intelligence models of emotion. In V. Hamilton, G. Bower & N. Frijda (Eds.), *Cognitive Perspectives of Emotion and Motivation* (pp. 287-320). London: Kluwer.

Picard, R.(1997). *Affective Computing*. Cambridge, MA: MIT Press.

Reisenzein, R. (2000a). Exploring the strength of association between the components of emotion syndromes: The case of surprise. *Cognition and Emotion*, *14*, 1-38.

Reisenzein, R. (2000b). The subjective experience of surprise. In H. Bless & J. Forgas (Eds.), *The message within: The role of subjective experience in social cognition and behavior* (pp. 262-279). Philadelphia, PA: Psychology Press.

Reisenzein, R. (1999). A theory of emotions as metarepresentational states of mind. *Personality and Social Psychology Review*. (Under review)

Reisenzein, R. (1996). Emotional action generation. In W. Battmann & S. Dutke (Eds.), *Processes of the molar regulation of behavior* (pp. 151-165). Lengerich: Pabst Science Publishers.

Russell, S., & Norvig, P. (1995). *Artificial Intelligence - A Modern Approach*. Englewood Cliffs, NJ: Prentice Hall.

Shafer, G., & Pearl, J. (Eds.) (1990). *Readings in Uncertain Reasoning*. Palo Alto, CA: Morgan Kaufmann.

Schützwohl, A., & Reisenzein, R. (1999). Children's and adults' reactions to a schema-discrepant event: A developmental analysis of surprise. *International Journal of Behavioral Development*, *23*, 37-62.

Stiensmeier-Pelster, J., Martini, A., & Reisenzein, R. (1995). The role of surprise in the attribution process. *Cognition and Emotion*, *9*, 5-31.

Modeling the Interplay of Emotions and Plans in Multi-Agent Simulations

Stacy Marsella (marsella@isi.edu)
USC Information Sciences Institute, 4676 Admiralty Way
Marina del Rey, CA 90292 USA

Jonathan Gratch (gratch@ict.usc.edu)
USC Institute for Creative Technologies, 13274 Fiji Way
Marina del Rey, CA 90292 USA

Abstract

The goal of this research is to create general computational models of the interplay between affect, cognition and behavior. These models are being designed to support characters that act in virtual environments, make decisions, but whose behavior also suggests an underlying emotional current. We attempt to capture both the cognitive and behavioral aspects of emotion, circumscribed to the role emotions play in the performance of concrete physical tasks. We address how emotions arise from an evaluation of the relationship between environmental events and an agent's plans and goals, as well as the impact of emotions on behavior, in particular the impact on the physical expressions of emotional state through suitable choice of gestures and body language. The approach is illustrated within a virtual reality training environment.

Introduction

Emotions play a central role in our lives. A wealth of empirical research has revealed a complex interplay between emotions, cognition and behavior. Emotional state may impact decision-making, actions, memory, attention, voluntary muscles, etc., which, conversely, may influence emotional state (e.g., see Berkowitz, 2000). Teasing apart and understanding these complex relationships is not an easy undertaking.

Not surprisingly, given this complexity, there are also a wealth of emotional models, with starkly differing views concerning the relation between cognition and emotion. While some theories have argued that cognition has a central role in evoking emotions (Lazarus, 1991), others have argued for a more minor role (Zajonc, 1984). With regards to the effects of emotions, theories of emotion have historically posited them as a problem for cognition, an impediment to effective cognitive function. On the other hand, more modern theories view emotions as more helpful than problematic, for example, a mechanism that facilitates human adaptation (e.g. Lazarus 1991, Simon, 1967).

We come to this conundrum from a certain perspective. The focus of our work is on general software agents that model human performance in rich simulated worlds. In particular, we focus on virtual training environments where intelligent agents interact with a human participant to facilitate the training objectives.

Emotions play an important role in such environments by enhancing believability and realism, increasing a sense of empathy and attachment to synthetic characters, and adding to the suspense of the simulation. For example, one of our environments, Carmen's Bright IDEAS, is designed to teach mothers of pediatric cancer patients better problem solving skills (Marsella et al., 2000). The mother learns by interacting with agents in a simulated world that mirrors her own. In particular, emotional models are used to help the mother identify with a human-like agent who faces various social problems due to her child's cancer. Another example is the Mission Rehearsal Exercise, a training environment designed to teach decision-making skills in highly evocative situations (Swartout, et al., 2001). The system provides an immersive learning environment where participants can experience the sights, sounds and circumstances they will encounter in real-world scenarios while performing mission-oriented training (Figure 1). Emotional models are used to enhance the intensity of the experience by creating characters that can respond emotionally to the student's decisions.

These simulations are set apart by the complexity of the environments and, more importantly, the detailed cognitive, emotional and behavioral modeling required. The agents face a variety of social and physical challenges, requiring the generation and execution of complex multi-agent plans. Overall, this complexity distinguishes this effort from more abstract simulation environments designed to study long term interactions of simpler agents (e.g., Nicholson et al., 1998) or believable, non-human agents in games (Neal Reilly, 1996).

Although complex, these realistic simulation environments offer a unique opportunity to explore and evaluate issues that arise by virtue of the complexity and fidelity of the modeling. For example, the agents must be able to generate complex plans with multiple goals and sub-goals. These plans may need to evolve or be replaced over time. Therefore, a key issue arises as to how the dynamics of this process and the structure of

Figure 1: A scene from the Mission Rehearsal Exercise

the resulting plan relate to overall emotional state and its dynamics. Another key issue concerns the agents' behaviors. They must interact with human participants across a range of modalities in a way that appropriately conveys their underlying emotional state. The wide repertoire of human nonverbal behaviors must be modeled, both subtle and extreme behaviors, consistent with emotional state. Fundamental questions arise as to what behaviors are exhibited and how various cognitive and emotional factors mediate between alternative behaviors. Finally, the realism of these simulations affords a unique, albeit weak, form of evaluation. The realism here supports more direct comparison with human behavior under matching conditions.

In essence, we are suggesting that it can be useful to attack the emotion conundrum head on via comprehensive, realistic simulations. Such simulations raise interesting research questions for cognitive science. Indeed the relation is synergistic since research on human cognition and emotion drives the design of our models.

In this paper, we demonstrate how some of the daunting subtlety in human behavior can be modeled by intelligent agents, from the perception of events in the world, to the appraisal of their emotional significance, through to their outward impact on agent behavior. We put forth a domain-independent solution that focuses on the problem of modeling "task-oriented" emotions – emotions that arise from performance of a concrete task. We then go on to illustrate the application of this model to virtual training environments.

Plans, Emotion & Behavior

The agents we design must provide convincing portrayals of humans facing difficult, dangerous problems. In particular, they must exhibit emotionally revealing nonverbal behaviors and expressions consistent with deeply evocative/disturbing situations. These behaviors must also change in concert with the emotional state of the agents; obviously people express themselves differently when sad, happy or angry.

Of course, one cannot realistically convey emotions without realistically modeling the genesis of those emotions. Because planning is central to our agent's behavior, we first needed to address how agents' plans/goals lead to their emotions. Then, we needed to address the impact of emotion on behavior. The driving force behind our modeling efforts was psychological research on the relation of cognition, emotion and behavior. However, the development of the models also raised significant research issues.

Plans and Emotional Appraisal

Many psychological theories of emotion emphasize the tight relationship between emotions and cognition. Emotions clearly influence our decision-making (Clore et al., 1994; Fiedler & Bless, 2000). What is less recognized is the strong influence cognition has over emotion. For example, the same event could evoke a variety of emotional responses depending on our mental state: getting a flat tire could evoke anger or joy depending on if we want to reach or avoid our destination. Such events derive their emotion charge, not from some intrinsic emotion evoking properties, but from our interpretation of their significance. Much of the recent theorizing on emotion builds on this observation, arguing that emotions arise from a cognitive appraisal of how events impact our plans and goals (Ortony et al, 1988; Lazarus, 1991).

Such psychological findings are problematic for building realistic models of human emotion. Just as fans of different teams will respond differently to the score of a goal, intelligent agents must respond differently to events in the simulation, and in a way that appears coherent to a human observer. For an agent developer, however, psychological findings and theories are seldom cast in a way that easily translates to general computational models.

Fortunately, there has been a nice convergence between cognitive appraisal models of emotion and the technologies underlying intelligent agents. Thus, while appraisal theories are vague on how events relate to goals, artificial intelligence planning methods now provide elaborate "mental" structures and inference techniques to assess this relationship (see Weld, 1999). While cognition cannot be reduced merely to planning, such algorithms can provide a cornerstone for making appraisal theories more concrete. By maintaining an explicit representation of an agent's plans, they can easily reason about future possible outcomes – a key requirement for handling emotions like hope and fear that involve future expectations. Planning techniques also detect interactions between plans, for example, as

when the plans of one agent are incompatible with those of another – a key requirement for handling emotions like anger or reproach which typically involve multiple actors.

Modern planning techniques also support a rich model of how cognition influences one's emotional state. We can model some of the dynamic ebb and flow of human emotion by relating emotional appraisals to the current state of plans in an agent's memory. As plans grow and change through the planning process, so too the emotional state will change as a reflection of this process – in a sense providing a window into an agent's mental processes.

Finally, by providing an explicit and rich reasoning infrastructure, plan-based approaches facilitate models of how emotions impact decision-making. Emotional state can act as search control, focusing cognitive resources on specific goals or threats. It can also alter the overall character of problem solving. For example, negative emotions seem to lead to narrow focused problem solving while positive emotions lead to broader problem solving that attempts to achieve multiple goals simultaneously (Sloman, 1987).

Emotional State and Physical Behavior

Psychological research on emotion reveals its pervasive impact on physical behavior such as facial expressions, gaze and gestures (Argyle & Cook, 1976; Ekman & Friesen, 1969, 1971). These behaviors communicate considerable information about an individual's emotional state. This may be intentional, as in shaking a fist. On the other hand, behaviors such as rubbing one's thigh, averting gaze and raised eyebrows may have no explicitly intended role in communication, but they suggest considerable information about emotional arousal, attitudes and attention. Indeed, observers can reliably infer a person's emotions and attitudes from nonverbal behaviors (Ekman & Friesen, 1969). For example, depressed individuals may avert gaze and direct gestures inward towards their bodies. An angry person's nonverbal behavior tends, if unsuppressed, to align itself with the object of the anger (e.g., by confrontational stares or obvious avoidance of eye contact).

Such movements also serve to mediate the information available to the individual. For example, if a depressed individual's head is lowered, this also regulates the information available to the individual. Orienting on an object of fear or anger brings the object to the focus of perceptual mechanisms, which may have indirect influences on cognition and cognitive appraisal by influencing the content of working memory. Even a soothing behavior like rubbing an arm may serve to manage what a person attends to (Freedman, 1972).

These findings provide a wealth of data to inform agent design but such sources are descriptive, not proscriptive, often leaving open many details as to how alternative behaviors are mediated. Contemporary agent technology allows one to create rich physical bodies for intelligent characters with many degrees of physical movement. This forces one to directly confront the problem of emotional consistency. For example, an "emotionally depressed" agent might avert gaze, be inattentive, perhaps hug themselves. However, if in subsequent dialog the agent used strong communicative gestures such as beats (McNeill, 1992), then the behavior might not "read" correctly. Similarly, people don t tend to nonchalantly use deictic gesture while simultaneously averting their gaze due to mild feelings of anger or guilt. Such behavior may look un-natural, inconsistent, or may convey a different shade of meaning depending on context. Which is not to say that the overall mix of behaviors should always be monolithic. People do say one thing while expressing another. At the least, the mix of nonverbal behaviors often shade the meaning of what is said or communicated nonverbally. Returning to the previous example, if an agent does combine deictic gesture with gaze aversion, it may shade the interpretation dramatically, towards an expression of extreme emotion and a desire to control that emotion. For example, the agent is so disgusted with the "listener", they can t bear to look at them .

Implicit in these various concerns is that the agent has what amounts to a resource allocation problem . The agent has limited physical assets, e.g., two hands, one body, etc. At any point in time, the agent must allocate these assets according to a variety of demands, such as performing a task, communicating, or emotionally soothing themselves. For instance, the agent's dialog may be suggestive of a specific gesture for the agent's arms and hands while the emotional state is suggestive of another. The agent must mediate between these alternative demands in a fashion consistent with their goals and their emotional state.

Implementation

Implementations demand compromise. In our work we limit the scope of models by what agent technology currently does well, rather than trying to develop comprehensive but less general solutions. Thus, we focus on emotions arising from plan generation and execution, and ignore a number of potential sources of emotion, such as ego conflict. Similarly we focus on physical behavior, expressing emotion through body gestures and facial expressions, ignoring the myriad ways people communicate emotion through speech (and instead rely on pre-recorded voice clips for verbal communication).

An agent consists of three main components. The planner/executor maintains a representation of the world state, and develops, executes and repairs plans that achieve the agent's goals. STEVE (Rickel & Johnson, 1998) plays the role of the planner/executor in the

application described below, but variety of AI planning methods could serve this role. The other components implement the cognitive appraisal of emotions and manage their physical manifestation.

Cognitive Appraisal

As we alluded above, we focus on cognitive appraisals as they relate to an agent's plans and draw on the strengths of modern artificial intelligence planning techniques. Specifically, we build on Émile, a computational realization of Ortony et al.'s cognitive appraisal theory (Gratch, 2000). The approach assesses the relationship between events and an agent's disposition (described by its goals, social standards). Unlike most computational accounts, Émile explicitly considers the role plans play in mediating the relationship between events and the agent's disposition. Rather than appraising events directly, Émile appraises the state of plans in memory, as inferred and elaborated by a general-purpose planning algorithm. This allows Émile to avoid the large number of domain-specific appraisal rules needed by prior computational approaches (e.g., Elliott, 1992). Domain-specific information, for the most part, is restricted to the operator descriptions (the domain theory) from which plans are built, and which an intelligent agent needs anyway to inform planning and action selection.

Émile also draws heavily on the explicit plan representation to derive the intensity of emotional response. Émile incorporates the view of Oatley and Johnson-Laird (1987) and Neal Reilly (1996) that emotions are related to changes in the perceived probability of goal attainment. Intensity is broken down into the probability of the event in question (e.g. the probability of goal achievement or the probability of a threat) and the importance (utility) of the event to the agent, both of which are derived from the current plan structure. As intensity is based on the current plans, the assessment is a reflection of their current state and changes with further planning. Individual assessments are aggregated into a set of "leaky buckets" associated with each emotion, where these buckets represent the current intensity of different emotions.

Physical Focus

The key challenge of the behavior component is to manage the flexibility in an agent's physical presence in a way that conveys a consistent emotional state. Agents are represented by rich bodies with fully articulated limbs, facial expressions, and sensory apparatus. The implementation must control the degrees of freedom provided by the agent's body in a way that satisfies the constraints imposed by psychological findings

To address this problem we rely on the Physical Focus model (Marsella et al. 2000), a computational technique inspired by work on nonverbal behavior in clinical settings (Freedman, 1972) and Lazarus's (1991) delineation of emotion-directed versus problem-directed coping strategies. The Physical Focus model bases an agent's physical behavior in terms of what the character attends to, how they relate to themselves and the world around them, specifically whether they are focusing on themselves and thereby withdrawing from the world or whether they are focusing on the world, engaging it.

The model organizes possible behaviors around a set of modes. Behaviors can be initiated via requests from the planner/executor or started spontaneously when the body is not otherwise engaged. At any point in time, the agent will be in a unique mode based on the current emotional state. This mode predisposes the agent to use particular nonverbal behavior in a particular fashion. Each behavior available to an agent is categorized according to which subset of these modes it is consistent with. Any specific nonverbal behavior, such as a particular nod of the head, may exist in more than one mode and conversely a type of behavior, such as head nods in general, may be realized differently in different modes. Transitions between modes are based on emotional state.

Modes also influence an agent's sensitivity to external stimuli, currently in a simplistic fashion. Rather than modeling the full flexibility of how people can focus their perception and attention (Wells & Matthews, 1994), we provide a domain specific mechanism for ranking stimuli by their intensity and filtering certain stimuli depending on if the focus mode is inner or outer directed.

Grouping behaviors into modes attempts to mediate competing demands on an agents physical resources, especially gesturing and gaze, in a fashion consistent with emotional state. This grouping model is designed with the intent that it be general across agents. However, realism also requires that specific behaviors within each mode incorporate individual differences, as in human behavior. For example, we would not expect a mother's repertoire of gestures to be identical to that of an army sergeant.

In the current work, we model three modes of physical focus: body-focus, transitional and communicative (as opposed to the five modes discussed in Marsella et al., 2000). Body focus is marked by a self-focused attention, away from the conversation and the problem-solving behavior. Emotionally, it is associated with considerable depression or guilt. Physically, it is associated with the tendencies of gaze aversion, paused or inhibited verbal activity and hand to body stimulation that is either soothing (e.g., rhythmic stroking of forearm) or self-punitive (e.g., squeezing or scratching of forearm). The agent exhibits minimal communicative gestures such as deictic or beat gestures (McNeil 1992,

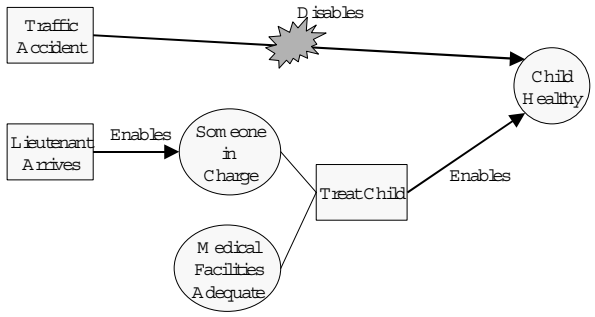


Figure 2: An example of the mother's plan

Cassell & Stone 1999) when in this mode. Transitional indicates an even less divided attention, less depression, a burgeoning willingness to take part in the conversation, milder conflicts with the problem solving and a closer relation to the listener. Physically, it is marked by hand to hand gestures (such as rubbing hands or hand fidgetiness) and hand to object gestures, such as playing with a pen. There are more communicative gestures in this mode but they are still muted or stilted. Finally, communicative indicates a full willingness to engage in the dialog and problem solving. Physically, it is marked by the agent's full range of communicative gestures, use of gaze in turn taking, etc.

Illustration

We illustrate the model by walking through an example of the system's behavior in the context of a virtual environment for familiarizing soldiers with the demands of peacekeeping operations. The Mission Rehearsal Exercise attempts to create an immersive learning environment through the integration of high-fidelity real-time graphics, intelligent agents, 3D audio and an interactive story whose outcome depends on the decisions and actions that participants take during the simulation.

In our working scenario, the system models a mix of three interactive and about forty pre-scripted virtual humans that play the parts of characters in the peacekeeping exercise. A human trainee commands a platoon of soldiers that have become involved in an automobile accident while driving to meet another platoon in need of reinforcement. The student must decide how best to allocate his forces between the conflicting goals of assisting an injured child and completing his mission, all under the watchful eyes of a "ZNN" cameraman.

Currently, only the character portraying the injured child's mother incorporates our emotional model. Figure 2 illustrates a simplified representation of the mother's plan at the opening scene in the scenario. The mother is waiting for the lieutenant (the student) to arrive, which she views as a precondition for her child to be treated. She is somewhat angry with the lieutenant, perceiving him as responsible for the accident (the domain-theory hard-codes an attribution that the lieutenant is responsible for "accident" task). This appraisal is moderated by the importance of the goal (high) and the likelihood of the threat cannot be overcome (moderate). Initially she believes the medical facilities are adequate to treat the child on scene, meaning she has the simple plan in memory that the lieutenant should arrive and her child will be treated, neither task being under her direct control. Since her child is hurt, a threat to an important goal, she has high levels of distress. The likelihood the treatment will be successful even if applied is relatively low (implying that there are many non-specific threats to its success) so she is also extremely anxious. The sense of hopelessness (and anxiety) leads her to have an inner-directed Physical Focus. Her body gestures are directed inward and she will not attend to most stimuli.

When the lieutenant arrives, the mother perceives that the sub-goal that someone is in charge is now attained and all non-specific threats associated with its attainment disappear. The probability that the child will be treated grows, and the mother's distress diminishes enough to transition her into transitional focus. Her gestures become more outward directed and she attends to more perceptual stimuli and her child.

Later in the scenario, the lieutenant orders one or two squads forward to reinforce the platoon downtown. The mother interprets this as disabling her sub-goal that the troops help her child. The strength of this interpretation is influenced by the number of squads the student orders forward (implemented by domain-specific rules that infer the probability of the disablement based on the number of moving units). The appraisal model treats this as a blameworthy event, causing the mother to become angrier at the troops. This anger is sufficient to transition her into communicative mode. The planner repairs the mother's current plan, deciding that imploring the troops to stay is a way of redirecting their behavior. Her body language in performing this action is colored by her body focus and anger level, either remaining seated and gesturing mildly or raising to a standing position and gesturing strongly (see Figure 1).

Discussion

This project is still in its early stages (the initial prototype was completed at the end of September 2000). From a research perspective the biggest limitation is the lack of evaluation. Is it a viable learning environment? Does the addition of emotional models increase the realism of the scenario? Do people find the character's reactions plausible? How do emotional models impact the learning experience? Our plan is to begin formal evaluations in the coming year in conjunction with other research groups in the psychology and communications departments at the University of Southern California. Our anecdotal feedback has been encouraging.

We have demonstrated the system to a number of military personal and those who served in Bosnia or Kosova seemed strongly affected by the experience. One U.S. Army Colonel began relating a related incident after seeing the demo, became quite emotional, and concluded by saying, "this system makes people feel, and we need that." In another anecdote, someone playing the role of the lieutenant became agitated when the mother character began yelling at him and when she wouldn't respond to his reassurances (she cannot be mollified when her anger exceeds some threshold).

Finally, there are a number of limitations in how the system infers emotional state that need adjustment or re-thinking in light of this application. As mentioned, cognitive appraisal only addresses emotions that arise from a concrete representation of plans of goals. We only weakly address the influence of emotion on perception and completely ignore the influence emotions hold over beliefs. Another key issue is the notion of responsibility. For example, whom should the mother blame for the accident? The troops? Herself? Our sense is she should have a shared sense of responsibility and that this sense should change dynamically, influenced by her emotional state and subsequent actions of the troops. Our treatment of anger is also too simplistic. Anger seems influenced by the extent to which we decide someone intended the offending action and the extent to which they show remorse or attempt to redress the offence. We suspect the explicit use of plans can assist in forming such assessments, but we are still sorting out how.

These limitations notwithstanding, the integration of plan-based appraisal of emotional state with the Physical Focus model provides a great deal of architectural support for emotional modeling. Furthermore, anecdotal evidence suggests that people find the agent's emotions to be plausible, and, to our surprise, people occasionally responded emotionally to our agents.

References

Argyle, M., & Cook, M. (1976) Gaze and mutual gaze. Cambridge University Press.

Berkowitz, L. (2000). Causes and Consequences of Feelings. Cambridge University Press..

Cassell, J. & Stone, M. (1999). Living Hand to Mouth: Psychological Theories about Speech and Gesture in Interactive Dialogue Systems. AAAI Fall Symposium on Narrative Intelligence.

Clore, G., Schwarz, N., & Conway, M. (1994). Cognitive causes and consequences of emotion. In Wyer & Srull (eds.), Handbook of social cognition, 2nd ed.

Ekman, P. and Friesen, W.V. (1971). Constants across cultures in the face and emotion. Personality and Social Psychology, 17 (2): 124-129.

Ekman, P. and Friesen, W.V. (1969). The Repertoire of NonVerbal Behavior: Categories, Origins, Usage and Coding. Semiotica 1:49-97.

Elliott C.D. (1992). The Affective Reasoner: A Process Model of Emotions in a Multi-agent System. PhD Thesis (TR#32), Northwestern University.

Freedman, N. (1972). The analysis of movement behavior during clinical interview. In Studies in Dyadic Communcation, 153-175.

Fiedler, K. & Bless, H. (2000). The interface of affective and cognitive processes. In Frijda, Manstead & Bem (eds.), Emotions and Beliefs. Cambridge University Press.

Gratch, J., (2000). Emile: Marshalling passions in training and education. Proc. of the 4th International Conference on Autonomous Agents Barcelona, Spain.

Lazarus, R.S. (1991). Emotion and Adaptation. Oxford Press.

Marsella, S. Johnson, W.L. & LaBore, C. (2000). Interactive Pedagogical Drama. Proceedings of the Fourth International Conference on Autonomous Agents. Barcelona, Spain, 301-308.

McNeil, D. (1992). Hand and Mind. University of Chicago Press, Chicago IL..

Neal Reilly, W.S., (1996). Believable Social and Emotional Agents. PhD Thesis CMU-CS-96-138. Carnegie Melon University.

Nicholson, A.E., Zukerman, I. & Oliver, C.D. (1998). Towards a Society of Affect-driven Agents. In Proceedings of the 20th Cognitive Science Society, Madison, WI.

Oatley, K. & Johnson-Laird, P.N. (1987). Towards a Cognitive Theory of Emotions. Cognition and Emotion, 1 (1).

Ortony A., Clore, G.L., & Collins, A. (1988). The Cognitive Structure of Emotions. Cambridge University Press.

Rickel, J. & Johnson, L. (1998). Animated agents for procedural training in virtual reality: perception, cognition, and motor control. Applied Artificial Intelligence (13), 343-382.

Simon, H.A. (1967) Motivational and emotional control of cognition. Psychological Review, 74, 29-39.

Sloman, A. (1987). Motives, mechanisms and emotions. Cognition and Emotion, 1, pp 217-234.

Swartout, W., Hill, R., Gratch, J., Johnson, W.L., Kyriakakis, C., Labore, K., Lindheim, R., Marsella, D., Moore, B., Morie, J., Rickel, J., Thiebaux, M., Tuch, L., Whitney, R. (2001). Towards the Holodeck: Integrating Graphics, Sound, Character and Story, in Proceedings of the Fifth International Conference on Autonomous Agents, Montreal, CANADA.

Weld, D. (1999). Recent Advances in AI Planning. AI Magazine 20 (2): 93-123.

Wells, A., and Matthews, G. (1994). Attention and emotion: a clinical perspective. Lawrence Erlbaum.

Zajonc, R.B. (1984). On the primacy of affect. American Psychologist. Vol. 39, No. 2, pp 117-123.

Elementary School Children's Understanding of Experimental Error

Amy M. Masnick (masnick@andrew.cmu.edu)
David Klahr (klahr@cmu.edu)
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Understanding data variability and potential error sources is essential to a full understanding of experimental science. We propose a typology of error that considers not only the nature of the error, but also the phase in the experiment in which it occurs. We looked at second- and fourth-grade children's understanding of error, their use of evidence in guiding this understanding, and the role of context in reasoning about error. We found that children could name and recognize sources of error even when they were unable to design controlled experiments. Children used evidence to guide their reasoning, making predictions and drawing conclusions based on the design of their experiments. Children were also sensitive to the context of reasoning: they differentiated the role of random error in relative and absolute measurements. These findings suggest that children understand a wide variety of potential error sources several years before they have acquired the systematic procedures necessary to control and interpret such error

A student in a fourth grade science laboratory is attempting to determine the effect of different factors on how far a ball rolls down a ramp. The ramp is adjustable for length, height, and surface smoothness, and there are two types of balls. Distance is measured by counting the number of discrete "steps" the ball travels up a sloped staircase at the end of the ramp. The teacher's primary instructional goal is to have the child learn how to design unconfounded experiments, in which experimental contrasts differing on only one attribute support logically grounded causal inferences.

In an ideal world, the student could demonstrate an understanding of experimental design by setting up two ramps such that they differed only with respect to surface type, releasing two identical balls, and observing that the ball on the smoother ramp went faster.

But the world is not ideal, and many unanticipated, unknown, and unintended events could influence the outcome of this experiment. Perhaps the student does not fully understand the logic of unconfounded experimentation, and so varies more than one factor at a time. Perhaps one ball is pushed slightly at the start of its roll or hits the side of the ramp on the way down. Perhaps one ball rolls back a few steps at the end of the run and the true distance it rolled is lost. Perhaps none of these "obvious" things occur, but, upon repeating the experiment several times, the student discovers that the balls travel different distances on each replication. What is the student to conclude?

Each of these events can be considered as a different type of experimental error. Among philosophers of science the question of how to best classify different types of experimental error remains controversial. The "conventional" view is that there is a true value and an error term that is part of the observed measurement, and it is distinguishing the magnitude of each that is the difficult part (cf., Hon, 1989). Quite a different perspective – and the one adopted in our analysis – is a process-based view that recognizes the inevitability of errors and that classifies them according to when in the overall cycle of experimental investigation they occur. Hon (1989) has proposed such a taxonomy and used it to organize a wide range of historical cases in which error played an important role in the scientific discovery process. We have adapted his approach by combining it with an earlier, psychologically oriented, classification of error (Toth & Klahr, 1999) to produce a similar taxonomy with which to organize some of the psychological literature on children's understanding of error, and to motivate our own investigations. In our taxonomy, there are five relatively distinct stages to the experimentation process (design, set-up, execution, measurement, and analysis of results), and each stage is associated with a category of error.

Types of Error in Designing and Executing Experiments

Design error This type of error occurs during the earliest conceptual phase of an experiment when some variables not being tested are not controlled and a confounded design is produced. For example, if the goal is to determine the effect of different ramp surfaces, then an experiment that compares a high smooth ramp with a low rough ramp contains a design error. No matter what the outcome of the test, it will be unclear whether any differences in the distance the balls rolled are due to the different steepnesses or different surfaces. Design errors occur "in the head" rather than "in the world" because they result from cognitive failures: either from a failure to fully understand the logic of unconfounded contrasts, or from inadequate domain knowledge.

Execution error These errors occur when something not considered or planned for in the design influences the outcome. Execution error can be random (such that replications can average out its effects) or biased (such that the direction of influence is the same on repeated trials), and it may be obvious (such as hitting the side of the ramp) or unobserved (such as an imperfection in the ball).

Measurement error This type of error overlaps the set-up and the measurement phases, because measurement is involved in setting up the apparatus and calibrating instruments as well as in assessing outcomes.

Interpretation error Interpretation errors can occur at any phase in the experiment. If an error occurs in any of the phases and is not recognized as such, it can influence the interpretation. The analysis phase involves both statistical analysis and theoretical inference, both of which are subject to a wide variety of statistical and cognitive errors. Errors that involve ascribing effects when in fact there are none, or claiming a null effect when one actually exists fall into this category.

Children's Understanding of Experimental Error

Although there is only a small philosophical literature on experimental error, there is an even smaller literature on the psychology of experimental error, i.e., empirical investigations of how people understand and interpret various kinds of experimental error. In this section we briefly summarize what is known about children's understanding of error. The terminology and methods of investigation in these developmental studies are quite varied, making it difficult to compare their results, but we summarize them in terms of the classificatory scheme presented earlier.

Several of the studies of grade school children's error understanding have focused on how children reason about repeated measurements and data variability. Varelas (1997) looked at third and fourth grade children's reasoning about repeated measurements when they carried out experiments in groups. She found that most children expected some variability in measurements, though why they expected this variability was not always clear.

Avoiding an error in the interpretation phase involves assessing when an error is sizeable enough to affect conclusions. Schauble (1996) looked at fifth and sixth graders, and non-college adults. One difficulty many children (and some adults) had was in distinguishing variation due to errors in measuring the results from variation due to true differences between the conditions (i.e., between intended contrasts and measurement phase errors). When in doubt, participants tended to fall back on their prior theories. If they expected a variable to have an effect, they interpreted variability as a true effect; otherwise, they were more likely to interpret the variability as due to error.

Lubben and Millar (1996) found that some high school children still have considerable difficulty understanding data variability, at least in situations in which they are given the data but are not performing the experiments themselves.

Error is a difficult concept to understand, and it seems likely that there are several levels of understanding (Lubben & Millar, 1996). There is evidence that first and second grade children can recognize a good experiment, even when they cannot yet generate one (Sodian, Zaitchik, & Carey, 1991). This finding suggests that children who are unable to generate error-related reasons for data variability may still have a basic understanding of error and therefore be able to recognize error-based explanations as plausible.

Causal Reasoning and Error Understanding

One way to clarify this literature about error is to reconceptualize error as a subset of the more fundamental topic of causal reasoning. Whether we recognize it as error or not, error is always caused by something. As suggested by the taxonomy proposed here, the nature of that something may be different for each phase of the experimental cycle. From this perspective, before one can reason about error in science experiments, it is necessary to be able to reason about causes (Koslowski & Masnick, in press).

In the current study, we set out to explore several aspects of children's understanding of error: their understanding and recognition of different types of error; consistency of reasoning about experimental design and conclusions, the ability to differentiate between the importance of error when comparing relative and absolute measurements; and the use of theory and evidence in justifications for confidence.

We know that children often have difficulty designing controlled experiments (e.g., Chen & Klahr, 1999). However, it is unclear how much they do understand and whether they can reason consistently based on their incorrect designs. If when children design confounded experiments, they make predictions about the outcome based on variables other than the target variable, it would suggest that although they do not fully understand the goals of the experiment, they use background knowledge to reason correctly based on the experiment they have designed.

Another measure of understanding is children's confidence about conclusions. If they understand the importance of a good design, we would expect them to be more confident of conclusions based on the results of an unconfounded experiment than a confounded one. At a more sophisticated level, if children understand the role of random error, they may still not be completely confident of outcomes after just one or two runs of an unconfounded test. They may consider that there are often uncontrollable factors that can affect a result and, by extension, the conclusions drawn.

The question of when error is important enough to alter conclusions is a difficult one. When looking at simple mechanics problems, the question also depends on the experimental context. If the goal is to determine the exact distance a ball will roll down a ramp under certain conditions, even the slightest unintended intrusion can raise questions about the result. But when the goal is to compare the relative distance a ball rolls given two levels of a particular variable, if the difference is sizeable, error is less important.

We designed a study to address several aspects of children's understanding of error throughout all phases of an experiment, to examine what elementary school children know about different types of error. First, we looked at whether children can design unconfounded experiments, make predictions consistent with their designs, and differentiate the role of error in absolute and relative measurements? Next, we looked at whether children generate alternative reasons for variation in repeated measurements, considering the roles of execution and measurement errors. Finally, we looked at whether children recognize potential sources of error.

Method

Participants Participants were 29 second-grade (mean age = 8.1) and 20 fourth-grade (mean age = 10.1) children from a private elementary school in southwestern Pennsylvania.

Materials Materials included two wooden ramps, each with an adjustable downhill track connected at its lower end to a slightly uphill, "staircase" surface. Children could set three binary variables to configure each ramp: the height (high or low), by using wooden blocks that fit under the ramps; the surface (rough or smooth), by placing inserts on the downhill tracks; and the length of the downhill ramp (long or short), by placing gates at either of two starting positions. Finally, children could choose either a rubber ball or a golf ball to roll down either ramp.

In addition, a laminated copy of a scale for indicating confidence (see below) and a stopwatch were used.

Procedure

Children were interviewed individually. All interviews were videotaped for later coding and analysis. During a brief familiarization phase, children were introduced to the ramp materials and the confidence scale (called a "sureness" scale, with levels of totally sure, pretty sure, kind of sure, and not so sure) and were asked a few questions to ensure that they understood how to use all of the materials.

Part One The purpose of Part One was to determine the extent to which children could design unconfounded experiments with these materials, and to assess their ability to differentiate between absolute and relative measurements. Each child was asked to design four experiments to determine the effect of different settings for specific variables that might affect how far a ball rolls down a ramp. In the first two experiments, children were asked to set up the ramps to test whether the steepness of the ramp made a difference in the outcome; in the third and fourth experiments, they were asked to test the effect of surface.

After the child set up the ramps, the experimenter asked why the ramps had been set up that way and also asked which ball was expected to go farther and why. Next, the experimenter asked the child to release both gates at the same time to see how far the balls rolled.

After the balls had stopped rolling, the experimenter asked the child what he/she had learned and why. Next, the experimenter asked the child whether the target variable (steepness or surface) made a difference. The child used the sureness scale to indicate confidence that the particular variable did make a difference, and to explain why.

Next, the experimenter asked a series of questions about relative and absolute values of the outcome variable. The experimenter asked the child to imagine, if the identical experiment were to be repeated, whether the same ball would go farther, and then whether the two balls would be expected to land on the exact same steps. The children were then asked to rate their sureness about each answer and explain why.

After each experiment and question series, the ramps were disassembled and the child was asked to set up the next experiment.

Part Two The purpose of this part was to explore the child's understanding of data variability in replicated experiments. A single ramp was set up with a high steepness, smooth surface, long run, and a golf ball. For each of five trials, the child was instructed to release the ball by lifting the gate on the experimenter's signal, while the experimenter simultaneously started a stopwatch. When the ball reached the bottom of the ramp (but before it began to roll up the steps), the experimenter stopped the stopwatch and read out a time for the child to record by writing it down. To ensure that all children were presented with the same range of data, the experimenter reported a fixed, predetermined set of times to each child, regardless of the actual time on the stopwatch[1].

At the completion of the five trials, the experimenter noted that it appeared to take a different amount of time for each roll and asked the child to generate reasons to explain these differences. Each child was encouraged to give as many reasons as he or she could think of, and then was asked for a summary explanation he or she would provide the teacher if she asked how long it took.

The experimenter then changed the surface of the ramp to a rough surface, and the ball was again rolled down five times. Again, the experimenter read each child an identical list of run durations. In this second round of numbers, there was a noticeable outlier among the numbers given[2]. After the five trials were completed, the experimenter again asked the child for reasons why the numbers would come out differently and a summary for the teacher.

Part Three Whereas Part Two required children to generate potential sources of error, in Part Three a few such sources were provided, to see how well children

[1] Times: 1.08, 1.20. 1.15, 1.02, 1.17; mean = 1.12, sd = 0.07

[2] Times: 1.90, 2.48, 1.88, 1.95, 1.85; mean = 2.01, sd = 0.26

could reason about their possible influence. Questions about both relative and absolute differences were asked.

The experimenter explained that she had been working with some children at another school who were trying to figure out whether run length made a difference. She demonstrated their situation by presenting the two ramps set up as an unconfounded experiment comparing the short and long run length, with both ramps having high steepness, smooth surfaces, and rubber balls.

The experimenter then asked about three scenarios the students had encountered: 1) one ball hit the side of the ramp on the way down; 2) the two balls were released at different times instead of simultaneously; 3) one ball rolled back a few steps before anyone could record how far it went. Note that scenarios 1 and 3 might be expected to affect both relative and absolute outcomes, while scenario 2 was designed as a control question because it should not effect on the outcome. For each, the experimenter asked the child whether the event described could affect how far the ball went, and whether it could change which ball went farther.

Results

Experimental Design Skills

Children's experimental design skills were assessed by looking at the number of correct (unconfounded) experiments designed. A correct design contrasted the target variable while holding the other three variables constant. This assessed their ability to avoid design phase errors, and to demonstrate knowledge of the Control of Variables Strategy (CVS). We categorized two types of incorrect designs: confounded (contrast of the target factor and one or more other factors), or non-contrastive (no contrast of the target factor). There was a significant effect of grade on CVS performance, with second graders averaging 16% unconfounded experiments, and fourth graders averaging 40% ($t(47) = 2.89$; $p = 0.006$).

Predicting Experimental Outcomes

Children's predictions about which ball would go farther were coded as correct or incorrect. (If the two balls traveled the same number of steps, children were coded as predicting incorrectly, unless they predicted a tie.)

Overall, children were extremely good at predicting the outcomes of unconfounded experiments and significantly less accurate when predicting the outcomes of non-contrastive designs, as measured by Fisher's exact tests of association. [3] For all but the first experiment, there was a strong relationship between predictive accuracy and type of design (unconfounded, confounded, or non-contrastive). (See Table 1.) Children's predictions were most accurate when they designed unconfounded experiments, next most accurate when they designed confounded experiments, and least accurate when they designed non-contrastive experiments. The relationship was stronger for fourth graders than for second graders.

Table 1: Prediction accuracy, by type of experiment

Number Accurate	Uncon-founded	Con-founded	Non - contrastive
Expt. 1	10/10	28/32	5/7
Expt. 2**	9/9	20/27	5/13
Expt. 3*	12/18	16/21	3/10
Expt. 4**	14/14	14/21	6/14

*$p < 0.05$; **$p < 0.01$

Explanations for Predictions

Children's explanations for their predictions were coded for mention of the target variable (steepness or surface), non-target variables, and any prior outcomes.

To assess the consistency of responses, we examined the relationship between children's reasons for their predictions and the type of experiment they designed (unconfounded, confounded, or non-contrastive), using Fisher's exact tests of association. For all four experiments, there was a significant relationship between mention of the target variable and the type of experiment designed. Overall, children mentioned the target variable as an explanation for 92% of their unconfounded experiments, 61% of their confounded experiments and 14% of their non-contrastive experiments. When broken down by grade, there is still a significant relationship at both grade levels. Similarly, there was a significant relationship between type of experiment and mention of the non-target variable. Children said they based their prediction on one or more of the non-target variables 6% of the time when they designed unconfounded experiments, 56% of the time when they designed confounded experiments, and 61% of the time when they designed non-contrastive experiments.

Confidence

Children's responses to the questions, "Can you tell if X makes a difference?" were coded as yes/no responses. Children noted how sure they were of this answer by using the four-level confidence scale. Finally, children explained why they chose the sureness value they did.

Nearly all of the children were "kind of sure," "pretty sure," or "totally sure" about whether steepness makes a difference on the first two experiments (98% and 86%) and whether surface makes a difference on the third and fourth experiments (90% and 86%). Confidence was unrelated to whether the test was unconfounded, as assessed by Fisher's exact tests of association ($p > 0.10$ for each of the four experiments), but there is some evidence that it is related to the accuracy of prediction. Four Fisher's exact tests were performed to assess the relationship between accuracy of prediction and

[3] Six times children designed correct experiments but inaccurately predicted the outcome. These six comparisons all occurred during the third experiment, a comparison of the surfaces, in which either the two balls tied or the ball on the rough surface actually rolled farther than the ball on the smooth surface.

confidence in conclusions. The trend was significant in the third and fourth experiments. In the first experiment, not enough children were unsure to allow for a strong comparison. Children who correctly predicted the outcome were more likely to be sure than those who predicted incorrectly (See Table 2 for details).

Table 2: Percent sure of target variable's effect, by prediction accuracy

Percent sure	Correct prediction	Incorrect prediction	P-value
Expt. 1	98% (42/43)	100% (6/6)	1.000
Expt. 2	91% (31/34)	73% (11/15)	0.179
Expt. 3	97% (30/31)	78% (14/18)	0.054
Expt. 4	94% (32/34)	67% (10/15)	0.022

Comparing Relative/Absolute Replications

For each question about whether the same ball would go farther if the experiment were to be repeated, children first answered yes or no, and then rated their confidence. These two responses – yes/no and confidence level – were combined into a single 7-point ordinal variable, ranging from totally sure the same ball would not go farther to totally sure it would go farther, with not so sure as the midpoint. An analogous coding scheme was used for the questions about whether the balls would land in the exact same positions.

The reasons given for why children expected the same or a different outcome were coded for mention of any of a list of common responses, including the fact that nothing had changed, and the effect of the target variable (e.g., "This one is the steeper ramp so that will make it go farther").

When asked whether the same ball would go farther were the experiment repeated without changes, over 90% of the time children thought it would (i.e., they said that they were kind of sure, pretty sure, or totally sure that it would). This figure excludes cases in which the balls traveled the same distance. The expectations about whether the balls would land in the exact same position were more varied. About 50% of the time, children thought the two balls would not land in the same positions again, about 40% of the time children thought they would, and the remaining times they were unsure.

To test whether children had different expectations for replication of relative and absolute outcomes, scores from the 7-point confidence code for absolute replication were subtracted from the corresponding scores for relative replication. For each child, we computed the mean of this difference score over the four experiments. Children were significantly more confident that the same ball would go farther than that the two balls would land in exactly the same place (mean difference = 2.46; sd = 1.76; t (48) = 9.8; $p < 0.001$). A t-test indicated a marginally significant effect of grade (mean for 2nd grade = 2.1, 4th grade = 3.0, t (47) = 1.95, p = 0.057). Children had different ideas about the importance of variation in the data depending on whether the judgments were about relative or absolute measurements.

Children's reasons for confidence about relative replication were nearly evenly divided: 40% were evidence-based (e.g., "this ball went farther last time"), and 45% theory-based (e.g., "this one will go farther because it's on the steeper ramp").

Accounting for Variability in Replications

Children's explanations for why the timing was different for each of the five trials were coded for mention of several factors, such as the child releasing the gate before or after the experimenter said "go," the experimenter stopping or starting the stopwatch too early or late, or the ball hitting the side of the ramp. Coding agreement over ten participants ranged from 85-100% on each code.

The average number of error sources named on the two sets of trials was 1.48 by second-graders and 2.15 by fourth-graders, a significant difference (t(46) = 2.73; p = 0.009). General linear models examining whether ability to design unconfounded experiments (as assessed by CVS score) is related to ability to name error sources in this second part indicate no relationship once grade is controlled in the model (F (1, 45) = 1. 79, p = 0.19).

Hypothetical Scenarios

Responses to the questions about hypothetical scenarios were classified into one of three categories: (1) yes, with mechanism explanation; (2) yes, without mechanism explanation; (3) no.

All participants correctly said that the ball hitting the side and that the ball rolling back a few steps could influence how far a ball went and whether the same ball would go farther. Eighty-eight percent were able to offer a reason for the former, and 72% were able to offer a reason for the latter. For the question about whether the timing of gate release would affect the distances traveled, 68% said that it would not and 4% offered a plausible mechanism for why it might make a difference (e.g., the vibration of the ramp might be different when a ball is simultaneously rolling down a ramp right next to it).

Overall, 34% of children completely and accurately answered all 6 questions. Fifty percent of the fourth graders and 22% of the second graders answered all the questions correctly.

Generating Error Sources

At several points throughout the interview, children were asked to think of reasons why experiments did not or might not have the same results when repeated (e.g., explaining why the balls would not land in the same place or why a ball rolled down the same ramp five times took a different amount of time for each run). The responses were coded for mention of possible sources of error, such as the ball hitting the side, or wind blowing, or the gates being lifted different ways, as a reason for the variation in results. Eighty-eight percent of children were able to

name at least one source of error. All six children who did not name any sources of error were second-graders.

Discussion

We set out to examine children's knowledge of different error sources. Our results indicate that despite children's difficulty in designing unconfounded experiments, they do understand a lot about error and its importance.

As found in earlier studies (Chen & Klahr, 1999; Toth, et al., 2000), children frequently made design errors (that is., they had difficulty designing unconfounded experiments). However, children showed consistency in their reasoning by referring to their design in justifying their predictions, regardless of whether it was a good experiment.

This evidence of consistency in justifications and conclusions also indicates some causal reasoning abilities. In the majority of cases, children recognized the link between the design and the outcome by considering their design to determine which factors would affect the outcome, and which factors they could draw conclusions about with confidence.

In addition, children's prior theoretical knowledge guided their reasoning. They were more likely to be confident about their conclusions when the evidence matched their prior beliefs (their predictions), though they were still confident more often than not, regardless of their predictions. Children also justified most predictions based on expected effects of the target and non-target variables (though this reliance varied based on design).

Children also demonstrated some understanding of the role of error in interpretation. Even by the second grade, children differentiated the importance of error in different situations, recognizing that errors are much more likely to affect measurements of exact positions than of relative positions. This finding may suggest a nascent understanding of the difference between main effects and specific examples. Children's confidence that the relative ordering would remain the same suggests they expect main effects to be robust, whereas their lack of confidence in absolute outcomes remaining the same suggests their understanding of variability in each sample.

When reasoning about experiments with ramps, children can use several different kinds of information. They can use their domain knowledge, i.e., what they know of the mechanics of friction and gravity, and of other factors that might affect how a specific instrument works. They can also use any formal experimental knowledge they have about what kinds of factors make for a good experiment, such as how to avoid errors in all phases of the experiment. Domain-specific knowledge enables them to name potential sources of error that could affect the outcome, while domain-general knowledge about experimental design encourages them to search for these specific examples.

The fact that most children are able to name at least one possible source of random error indicates that they do have at least a rudimentary idea about how unpredictable and uncontrolled factors can influence an experiment's outcome. At the same time, the observation that most of these same children are not consistently able to design unconfounded experiments suggests that the understanding is not complete at this age. Knowledge about this gap in understanding can lay the foundation for future research about children's knowledge of science experimentation and about the most effective means to aid science educators teaching children these skills.

Acknowledgments

This work was supported in part by grants from NICHHD (HD25211) and the James S. McDonnell Foundation (96-37). We thank Anne Siegel, Jolene Watson, and three anonymous reviewers for comments on earlier drafts.

References

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, *70*, 1098-1120.

Hon, G. (1989). Towards a typology of experimental errors: An epistemological view. *Studies in History and Philosophy of Science*, *20*, 469-504.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasonin*g. Cambridge, MA: MIT Press.

Koslowski, B., & Masnick, A. (in press). The development of causal reasoning. In U. Goswami (Ed.) *Blackwell Handbook of Childhood Cognitive Development*. Oxford, England: Blackwell Press.

Kuhn, D., Amsel, E., & O'Loughlin, M. (1988). *The development of scientific thinking skills*. Orlando, FL: Academic Press.

Lubben, F., & Millar, R. (1996). Children's ideas about the reliability of experimental data. *International Journal of Science Education*, *18*, 955-968.

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology*, *32*, 102-119.

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development*, *62*, 753-766.

Toth, E. E., & Klahr D. (1999). "It's up to the ball": Children's difficulties in applying valid experimentation strategies in inquiry based science learning environments. Paper presented at the Annual Convention of the American Educational Research Association. Montreal, Canada.

Toth, E. E., Klahr, D., & Chen, Z. (2000). Bridging research and practice: A cognitively-based classroom intervention for teaching experimentation skills to elementary school children. *Cognition and Instruction 18*, 423-459.

Varelas, M. (1997). Third and fourth graders' conceptions of repeated trials and best representatives in science experiments. *Journal of Research in Science Teaching*, *34*, 853-872.

Interactive Models of Collaborative Communication

Michael Matessa (mmatessa@arc.nasa.gov)
NASA Ames Research Center, Mail Stop 262-4
Moffett Field, CA 94035 USA

Abstract

The collaborative nature of communication has been demonstrated by research on the increased efficiency (Hupet & Chantraine, 1992) and the adaptive behavior (Giles, Mulac, Bradac, & Johnson, 1987) of interacting pairs, but these two lines of research have never been explicitly related. This paper reports empirical results showing that adaptively matching word use can increase communication efficiency and also gives an ACT-R (Anderson & Lebiere, 1998) modeling account of the processes involved.

Efficient Communication

Imagine that two people have to communicate a number of times about abstract figures that are difficult to name. Typically, the pair will initially use a long referential phrase and with subsequent references shorten that phrase to one or two words (Clark & Wilkes-Gibbs, 1986; Krauss & Fussell, 1991; Krauss & Weinheimer, 1966). For example, in an experiment run by Krauss and Fussell (1991), a pair shown the figure in Figure 1 referred to it over five trials as

a Martini glass with legs on either side
Martini glass with the legs
Martini glass shaped thing
Martini glass
Martini

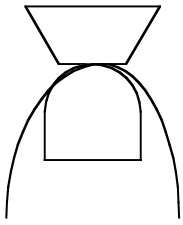


Figure 1: An abstract figure

This process is evidence of the collaborative nature of communication since subsequent phrases tend to make reference to previous phrases and since the phrase eventually agreed on to describe the object would not likely be able to describe the object without the benefit of the prior history of the evolution of the phrase.

Several partner-related factors have been shown to influence the number of words used in the referential communication task. If subjects are asked to create referential phrases for an imagined partner who will later read the phrases, the phrases tend not to decrease over time (Hupet & Chantraine, 1992). If a partner is present but not allowed to give feedback, the rate of decrease is slowed (Krauss & Weinheimer, 1966).

Accommodating Communication

In this discussion, accommodation is the matching of partner behavior in a conversational setting. These behaviors can include lexical choice (Fais, 1998; Garrod & Anderson, 1987; Garrod & Doherty, 1994) and syntactic choice (Bock, 1986; Fais, 1994) as well as speech styles, dialect, non-verbal behavior, vocal intensity, prosody, speech rate and duration and pause length (Giles, Mulac, Bradac, & Johnson, 1987).

Examples of accommodation can be seen in the maze game of Garrod and Doherty (1994), where subjects must decide how to describe their positions in a two-dimensional maze. Some subjects came to describe their positions in a line notation, giving first the line and then their location in that line:

A: Third row two along.
B: Second row three along.

Other subjects developed a matrix notation, giving horizontal and vertical locations:

A: Correct, I'm presently at C5.
B: E1.

Most research on accommodation has focused on dependent measures of converging/diverging behavior or recipient evaluations of that behavior. One hypothesis of this paper is that diverging behavior (non-accommodation) can not only influence the evaluation of behavior, but can also reduce the efficiency of referential communication. Support for this hypothesis would be shorter messages for subjects interacting with accommodating partners as compared to subjects interacting with non-accommodating partners. To do this manipulation with human partners, either confederate partners or partners motivated with positive and negative social group pressures would need to converge or diverge to communication behavior. Either choice would introduce extraneous social complications

into a question about informational processing. Ideally, the decision to diverge or converge should be independent of other communication processing in the partner. One solution is to use computational agents as partners. Two agents could be created that would either converge to or diverge from word choice of a human partner, with other communication processing being exactly the same. If both agents accommodated to the message length used by the human partner, then message length could be used as a dependent measure of efficiency. This would then test the effect of lexical accommodation on message length. The generality of the results would be greater if the behavior of the agents were psychologically plausible. One line of research involving computational theory of human cognition is ACT-R.

ACT-R & Communication

ACT-R (Anderson & Lebiere, 1998) is a computational theory of human cognition incorporating both declarative knowledge (e.g., addition facts) and procedural knowledge (e.g., the process of solving a multi-column addition problem) into a production system where procedural rules act on declarative chunks. At a subsymbolic level, facts have an activation attribute which influences their probability of retrieval and the time it takes to retrieve them. Rules have a reliability attribute which influences their probability of being used.

Support for this declarative/procedural viewpoint has been found in many ACT-R language projects. One project emphasizing declarative representation is Boyland and Anderson's (1997) model of syntactic priming. Research has shown that the use of a specific syntax can be primed in experimental settings if a subject repeats presented sentences (Bock, 1986). Boyland and Anderson created a model that explained this phenomenon as priming of declarative structures built from the comprehension of sentences.

With a procedural representation, Matessa and Anderson (2000) showed that the ACT-R rule reliability learning mechanism predicts a blocking effect in cue learning where the use of highly available cues can block the learning of more reliable cues, since the sequential nature of productions allows only one cue to be chosen at a time. This prediction was supported by experimental evidence of blocking for linguistic actor choice cues such as word order, case marking, and verb/noun matching. Taatgen and Anderson (2000) used a model that combined both declarative and procedural learning to explain the U-shaped learning of irregular verbs, and Lewis (1998) created a parsing model with retroactive and proactive interference of declarative knowledge and procedural attachment processes.

Any interactive model of communication must be able to establish mutual knowledge, interpret the communicative intent of a partner, follow basic communicative obligations, and use communication to further some goal. These abilities have been the focus of a number of lines of research in the communication literature (Clark & Schaefer, 1989; Core & Allen, 1997; Poesio & Traum, 1998; Traum & Allen, 1994) and the ACT-R model of communication presented in this paper is guided by theories in this literature. ACT-R itself is a method for describing human cognition in terms of facts and rules, but the content of the facts and rules used in communication must be guided by current theories of communication. This model of communication was used to test the effect of accommodation (the matching of partner vocabulary) on communication efficiency by having two ACT-R models created from the basic communication model, one accommodating to word use and one non-accommodating.

Experiment

Communication is usually motivated by the desire to complete a certain task. Subjects were given parts of a graph with the goal of creating a whole graph. The graphs are colored objects connect by lines (similar to those used by Levelt (1982) to study communicative reference) and are designed so that similarly colored objects on the parts can overlap and form a larger graph.

Communication using text is more conducive to modeling, so the subjects send messages by way of a chat window from two different computers. In addition to creating a whole graph from two parts, subjects also have the goal of confirming each of the circles. This is done by each subject selecting one circle at a time -- if the circles are the same, their score is increased, but if the circles are different, the score is decreased. This confirmation goal gives an objective measure of task performance in terms of a score, and it allows for the use of more complicated dialogue acts such as requesting that the other person confirm a circle or committing to confirming a circle.

In a similar spirit to the COLLAGEN project (Rich & Sidner, 1998), this modeling effort is not aimed towards the processing of unrestricted English syntax but in modeling the higher-level communicative acts accomplished with English. So like the COLLAGEN project the models interact with people with a restricted set of English phrases. This restricted interface need not drastically hinder the communication process or task performance. In a study comparing a restricted interface to an unrestricted interface for students solving physics problems, Baker and Lund (1997) showed that the restricted communication interface did not interfere with task performance. In fact, it promoted a more task-

focused and reflective interaction.[1] Still, for the current task, unrestricted and restricted communication were compared to see if the restricted interface had any effect on task performance. The restricted interface allows the composition of a text message by first choosing a topic of discussion and dialogue act to address the topic. The topics of conversation are paired connections (how one circle relates to another), multiple connections (rows or columns of circles), numbers (how many of a specific kind of circle there are), correspondences (what circle in one person s graph corresponds to in the other person s graph), confirmations (talking about mutually confirming a circle), and experiment phases.

Also, to allow more problems to be solved in a single experimental session which would allow the development of communication over time, the problems were simplified to have six total objects with one marked as common. From previous research (Clark & Wilkes-Gibbs, 1986; Krauss & Fussell, 1991; Krauss & Weinheimer, 1966) it was expected that the message length would decrease over time. To facilitate this decrease in the restricted interface, the manner of composing messages in the template was changed from choosing words from a pull-down menu to typing words that were displayed in a menu. The menu for the word choice could be skipped over with the Tab key, and in this way shorter messages could be produced. This new method permits a closer correspondence to the unrestricted interface (unrestricted typing) and gives a time benefit to skipping words by not having to spend time in typing them. Additional dimensions of size and shape were added to the color dimension of the circles in order to provide more redundant information in the problem that could later be left out of messages, resulting in a shorter message length. These dimensions were redundant, so that red objects were always small and thin, green objects were always medium and round, and blue objects were always large and fat.

Subjects

One hundred Carnegie Mellon University undergraduates attempted the graph completion task. Twenty-two were paired and used the unrestricted interface, thirty-two were paired and used the restricted interface, twenty-two were paired with an accommodating ACT-R model, and twenty-four were paired with a non-accommodating ACT-R model. This created eleven pairs in the unrestricted interface condition, sixteen pairs in the restricted interface condition, twenty-two pairs in the accommodating model condition (pairs consisting of a subject and a model), and twenty-four pairs in the non-accommodating model condition.

[1] For an alternate viewpoint, see Suchman (1997).

Method

Each pair was told that they would each be given part of a graph and their goal was first to create a whole graph as a result of circles overlapping from each part of the graph, and then to confirm each circle in the whole graph. They were told they would be sitting in different rooms and would be using a chat window to talk to each other. They were shown a drawing pad which contained an example graph part consisting of connected colored circles, and were shown how to add and erase circles representing circles from the partner s graph. They were also shown a chat window which could send eighty-character messages and only displayed the partner s last message. In the restricted interface condition, subjects were told that messages were composed in a communication window that allowed the creation of restricted sentences and were led through the creation of each kind of message. After making sure subjects understood the task, they were then given individual practice problems which used the adding, erasing, and confirming functions of the drawing pad. Finally, the subjects were given their graph parts and were told there were no time constraints in solving the problem.

Results

Figure 2 shows the average time that pairs in each condition took to solve problems. Error bars in this and subsequent figures represent standard error. Results are averaged for the first three problems, the second three, third, and fourth. Since there were an unequal number of pairs in each condition for any particular problem, statistics are performed on each group of three problems.

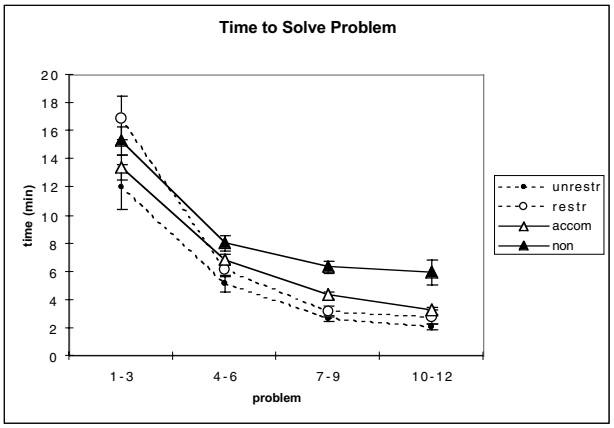


Figure 2: Time to solve problem

There was no significant effect of condition on time to solve problems in the first two groups of three problems ($F(3,65)=2.12$; $F(3,56)=1.72$), but there was an effect in the last two groups of three problems ($F(3,37)=8.58$, $p<.0005$; $F(3,28)=5.76$, $p<.005$). This effect is driven

by slower times in the non-accommodation condition, which was significantly slower than the accommodation condition in the last two groups of three problems ($t(24)=3.71$, $p<.001$; $t(18)=2.61$, $p<.05$). Since the results were similar between the restricted and unrestricted conditions for errors and time to solve problems, the unrestricted condition will not be included in subsequent discussions.

Figure 3 shows how many words were typed, or message length, for sentences concerning connections of objects. Message length tended to decrease with time, for example, a message such as The small thin red object is above our large fat blue object in the first problem could be reduced to messages such as red above blue by the twelfth problem.

Messages in the non-accommodation condition tended to be longer than those in the accommodation condition, not significantly in the first two groups of three problems ($t(43)=0.55$; $t(39)=0.83$) but significantly in the second two groups of three problems ($t(24)=1.97$, $p<.05$; $t(18)=1.81$, $p<.05$).

Human Model

The accommodating and non-accommodating models are able to solve the communication task, but cannot by themselves explain the effect of accommodation. This is because they are "passive" in that they are not the first to decide to skip words in messages descriptions or to skip messages describing confirmation actions. Instead, they follow the lead of their partner and skip words when their partner skips words and skip messages when their partner skips messages. What is needed is an "assertive" "human" model that can decide to skip words and messages first. This model should also be able to account for differences found when subjects interact with accommodating and non-accommodating models.

This "human" model was created by extending the accommodation model with extra rules for actively skipping words. Since time is saved by not typing, these rules make solving the problem more efficient. This effect is achieved in the current situation by having the efficiency rules be sensitive to cooperative actions of the partner (with accommodative word matching signaling cooperative behavior). Two of the rules, skip-word-match-eff and skip-confirm-match-eff, attempt to retrieve memories of their partner matching their own word use. This gives these rules a sensitivity to whether their partner is accommodating or non-accommodating. The other rules do not attempt to retrieve matching memories. The rationale behind these rules is that the decision to skip a word or confirmation message will more likely lead to success if the partner has been cooperative in their behavior, and memories of word matching by the partner give evidence of this cooperation. Rules that find this evidence have a higher reliability because the evidence increases the probability that skipping will lead to success. The rule to continue confirmation

These rules have a subsymbolic value, reliability, associated with them that affects the probability with which they will be used -- rules with higher reliabilities have a higher probability of being used. The efficiency rules added to the accommodation model were skip-word-match-eff (skips a nonessential word if partner matches word) with a reliability of .735 and skip-word-nomatch-eff (skips a nonessential word) with a reliability of .730. The reliability values were set with regard to subject performance with the accommodating and non-accommodating models. These values were then used in runs of the human model with another human model to make zero-parameter predictions of subject performance with other subjects in the restricted interface condition.

Looking at message length, Figure 3 shows results of twenty runs of the "human" model (shown as a dashed line) interacting with the accommodating model, the non-accommodating model, and another human model compared to the results of subjects interacting with the accommodating model, the non-accommodating model, and another human subject.

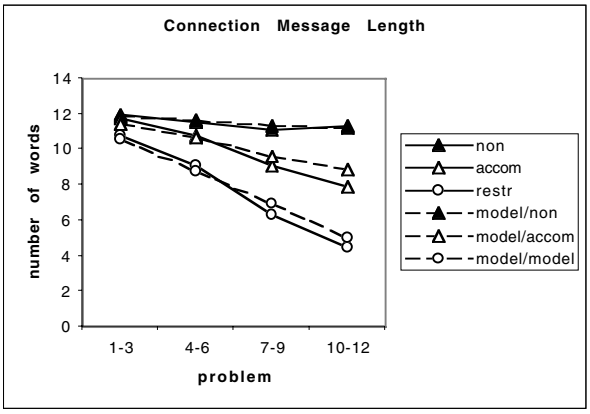


Figure 3: Connection message length

Conclusions

Data from the main experiment show that subjects interacting with accommodating models that match their word choice can solve problems faster than subjects interacting with non-accommodating partners. This result ties together results from the referential communication literature showing partner-based effects on efficiency with results from the accommodation literature showing accommodating behavior motivated by efficiency.

Having a theory of communication in a computational form allowed testing of the theory by having it directly interact with subjects. In terms of errors and time to solve problems, subjects generally

reacted to the accommodating model incorporating the theory much like any other human. In fact, in a post-experiment questionnaire subjects guessed they were interacting with a human 43% of the time they were interacting with the accommodating model (subjects guessed the non-accommodating model was human 48% of the time, but the difference is not significant). There is still room for improvement however, since only 10% of subjects interacting with human partners thought their partners were computers.

The computational nature of the theory also allowed predictions to be made without the use of human subjects. The reliability of the efficiency rules for the human model were set based on human performance with the accommodation and non-accommodating models, but the results of the human model communicating with another human model represent a zero-parameter prediction that closely matched human performance in the restricted interface condition.

Acknowledgments

This research was sponsored as part of the author s doctoral thesis by the Office of Naval Research under contract number N00014-95-10223 to John Anderson at Carnegie Mellon University.

References

Anderson, J. R., & Lebiere, C. (1998). *The Atomic Components of Thought*: Mahwah, NJ: Erlbaum.

Baker, M., & Lund, K. (1997). Promoting reflective interactions in a CSCL environment. *Journal of Computer Assisted Learning, 13(3)*, 175-193.

Bock, K. J. (1986). Syntactic persistence in language production. *Cognitive Psychology, 18*(3), 355-387.

Boyland, J. T., & Anderson, J. R. (1997). *Comprehension and production as avenues of syntactic priming.* Paper presented at the 19th Annual Conference of the Cognitive Science Society.

Budiu, R., & Anderson, J. R. (2000). *Integration of Background Knowledge in Sentence Processing: A Unified Theory of Metaphor Understanding, Semantic Illusions and Text Memory.* Paper presented at the 3rd International Conference on Cognitive Modeling, Groningen, Netherlands.

Clark, H., & Schaefer, E. (1989). Contributing to Discourse. *Cognitive Science, 13*, 259-294.

Clark, H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1-39.

Core, M. G., & Allen, J. F. (1997). *Coding dialogs with the DAMSL scheme.* Paper presented at the AAAI Fall Symposium on Communicative Action in Humans and Machines, Boston, MA.

Fais, L. (1994). Conversation as collaboration: Some syntactic evidence. *Speech Communication, 15*, 231-242.

Fais, L. (1998). Lexical accommodation in human- and machine-interpreted dialogues. *International Journal of Human-Computer Studies, 48*, 217-246.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition, 27*, 181-218.

Garrod, S., & Doherty, G. (1994). Conversation, co-ordination and convention: an empirical investigation of how groups establish linguistic conventions. *Cognition, 53*, 181-215.

Giles, H., Mulac, A., Bradac, J. J., & Johnson, P. (1987). Speech accommodation theory: The first decade and beyond. *Communication Yearbook, 10*.

Hupet, M., & Chantraine, Y. (1992). Changes in repeated references: Collaboration or repetition effects? *Journal of Psycholinguistic Research, 21*(6), 485-496.

Krauss, R. M., & Fussell, S. R. (1991). Constructing shared communicative environments. In L. B. Resnick & J. M. Levine (Eds.), *Perspectives on socially shared cognition* (pp. 172-200).

Krauss, R. M., & Weinheimer, S. (1966). Concurrent feedback, confirmation and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology, 4*, 343-346.

Levelt, W. J. M. (1982). Linearization in describing spatial networks. In S. Peters & E. Saarinen (Eds.), *Processes, beliefs, and questions* (pp. 199-220): Dordrecht: D. Reidel.

Lewis, R.L. (1998) Working Memory in Sentence Processing: Retroactive and Proactive Interference in Parsing. Talk presented at CUNY 98, New Brunswick, New Jersey.

Matessa, M., & Anderson, J. R. (2000). Modelling Focused Learning in Role Assignment. *Language and Cognitive Processes, 15*(3), 263-292.

Poesio, M., & Traum, D. (1998). Towards an Axiomatization of Dialogue Acts. In J. Hulstijn & A. Nijholt (Eds.), *Proceedings of the Twente Workshop on the Formal Semantics and Pragmatics of Dialogues (13th Twente Workshop on Language Technology)* (pp. 207-222). Enshede, NL.

Rich, C., & Sidner, C. L. (1998). COLLAGEN: A Collaboration Manager for Software Interface Agents. *User Modeling and User-Adapted Interaction, 8(3/4)*, 315-350.

Suchman, L. (1997). Do categories have politics? In B. Friedman (Ed.), Human values and the design of computer technology. Cambridge: Cambridge University Press.

Taatgen, N., & Anderson, J. R. (2000). *Why do children learn to say "broke"? A model of learning the past tense.* Paper presented at the Seventh Annual ACT-R Workshop, Pittsburgh, PA.

Traum, D., & Allen, J. (1994). *Discourse Obligations in Dialogue Processing.* Paper presented at the 32nd Annual Meeting of the Association for Computational Linguistics.

Testing the Distributional Hypothesis: The Influence of Context on Judgements of Semantic Similarity

Scott McDonald (scottm@cogsci.ed.ac.uk)
Michael Ramscar (michael@cogsci.ed.ac.uk)
Institute for Communicating and Collaborative Systems, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW Scotland

Abstract

Distributional information has recently been implicated as playing an important role in several aspects of language ability. Learning the meaning of a word is thought to be dependent, at least in part, on exposure to the word in its linguistic contexts of use. In two experiments, we manipulated subjects' contextual experience with marginally familiar and nonce words. Results showed that similarity judgements involving these words were affected by the distributional properties of the contexts in which they were read. The accrual of contextual experience was simulated in a semantic space model, by successively adding larger amounts of experience in the form of item-in-context exemplars sampled from the British National Corpus. The experiments and the simulation provide support for the role of distributional information in developing representations of word meaning.

The Distributional Hypothesis

The basic human ability of language understanding – making sense of another person's utterances – does not develop in isolation from the environment. There is a growing body of research suggesting that *distributional information* plays a more powerful role than previously thought in a number of aspects of language processing. The exploitation of statistical regularities in the linguistic environment has been put forward to explain how language learners accomplish tasks from segmenting speech to bootstrapping word meaning. For example, Saffran, Aslin and Newport (1996) have demonstrated that infants are highly sensitive to simple conditional probability statistics, indicating how the ability to segment the speech stream into words may be realised. Adults, when faced with the task of identifying the word boundaries in an artificial language, also appear able to readily exploit such statistics (Saffran, Newport & Aslin, 1996). Redington, Chater and Finch (1998) have proposed that distributional information may contribute to the acquisition of syntactic knowledge by children. Useful information about the similarities and differences in the meaning of words has also been shown to be present in simple distributional statistics (e.g., Landauer & Dumais, 1997; McDonald, 2000).

Based on the convergence of these recent studies into a cognitive role for distributional information in explaining language ability, we call the general principle under exploration the *Distributional Hypothesis*. The purpose of the present paper is to further test the distributional hypothesis, by examining the influence of context on similarity judgements involving marginally familiar and novel words. Our investigations are framed under the 'semantic space' approach to representing word meaning, to which we turn next.

Distributional Models of Word Meaning

The distributional hypothesis has provided the motivation for a class of objective statistical methods for representing meaning. Although the surge of interest in the approach arose in the fields of computational linguistics and information retrieval (e.g., Schutze, 1998; Grefenstette, 1994), where large-scale models of lexical semantics are crucial for tasks such as word sense disambiguation, high-dimensional 'semantic space' models are also useful tools for investigating how the brain represents the meaning of words.

Word meaning can be considered to vary along many dimensions; semantic space models attempt to capture this variation in a coherent way, by positioning words in a geometric space. How to determine what the crucial dimensions are has been a long-standing problem; a recent and fruitful approach to this issue has been to label the dimensions of semantic space with *words*. A word is located in the space according to the degree to which it *co-occurs* with each of the words labelling the dimensions of the space. Co-occurrence frequency information is extracted from a record of language experience – a large corpus of natural language. Using this approach, two words that tend to occur in similar linguistic contexts – that is, they are *distributionally* similar – will be positioned closer together in semantic space than two words which are not as distributionally similar. Such simple distributional knowledge has been implicated in a variety of language processing behaviours, such as lexical priming (e.g., Lowe & McDonald, 2000; Lund, Burgess & Atchley, 1995; McDonald & Lowe, 1998), synonym selection (Landauer & Dumais, 1997), retrieval in analogical reasoning (Ramscar & Yarlett, 2000) and judgements of semantic similarity (McDonald, 2000).

Contextual co-occurrence, the fundamental relationship underlying the success of the semantic space approach to representing word meaning, can be defined in a number of ways. Perhaps the simplest (and the approach taken in the majority of the studies cited above) is to define co-occurrence in terms of a 'context window': the co-occur-

rence frequency of w_1 with w_2 is defined as the number of times that w_2 (the 'context word') occurs in the window of n words surrounding w_1, summed over all instances of w_1 in the corpus. Given a set of k context words, any word in the vocabulary can be represented as a k-dimensional vector of co-occurrence frequencies. The best fit to psychological data is typically achieved with word vectors constructed using context window sizes between ±2 and ±10 words (see, e.g., Patel, Bullinaria & Levy, 1998).

Besides its emphasis on identifying a potential source of information useful for the development of semantic representations, the distributional hypothesis also accommodates predictions about the consequences of *manipulating* the learning environment. By modifying the degree of distributional similarity holding between two words in a person's language experience, a particular word's location in semantic space can be adjusted (i.e., a word vector can be 'pushed' in a given direction). In Experiments 1 and 2 we test whether manipulating contextual co-occurrence has behavioural consequences, by eliciting judgements of semantic similarity involving marginally familiar and nonce words embedded in biasing contexts.

Learning Word Meaning from Context

It is well-established that the context in which an unfamiliar word occurs is an important determinant of how much is learned about the word, and it is apparent that context often provides the sole means for establishing its meaning (e.g., Carnine, Kameenui & Coyle, 1994; Fischer, 1994). In order to interpret an unknown word, the context provides cues, in the form of some combination of: (1) the identity of the words in the context surrounding the unknown word and the relationships between these words and the unknown word (i.e., distributional information); (2) world knowledge retrieved from long-term memory associated with these words; and (3) the cognitive model of the discourse (or situation) currently being built. But it seems that distributional information on its own, if suitably constraining, could be sufficient for determining the meaning of an unfamiliar word. Consider the occurrence of the neologism *broamed* in the following context:

Because the capsule was hermetically broamed, its contents were in perfect condition after more than a hundred years under water.

In this example, knowledge about the distributional behaviour of *hermetically* certainly guides the inference that the meaning of *broamed* is similar to the meaning of *sealed*, because *hermetically* nearly always co-occurs with *sealed*. Further support for this inference is contributed by knowledge about capsules and the conditions required in order for something to remain in perfect condition in adverse circumstances.

Contextual cues also play an important role in consolidating the meaning of newly-learned words. The more exemplars of a word in its context of use that are encountered, the more its meaning can be refined and delimited, especially if one has some prior knowledge of the discourse or passage topic. We assume that a close correspondence exists between a word's subjective familiarity and the amount of experience one has with the word. The less experience, the less familiar the word and the less established its semantic representation in the brain.

In the experiments reported below, we attempt to manipulate the distributional knowledge associated with sets of marginally familiar and completely novel words in order to test a basic prediction of semantic space models in particular and the distributional hypothesis in general. Distributional information is the only variable manipulated; for each item we constructed two different paragraph contexts, each containing only four exemplars of the item. By judicious selection of the words in the context surrounding each instance of the word of interest, co-occurrence patterns can be created that resemble the patterns of other, more familiar words. Using semantic space model terminology, a word vector can be 'pushed' towards another vector by bringing dimensions of the space into alignment. The question we addressed was whether this manipulation of distributional information was sufficient to influence subjects' ratings of semantic similarity.

Experiment 1

Experiment 1 focuses on marginally familiar words. These are words that one is likely to have encountered, but not with sufficient frequency to have a firm grasp of their meaning. For instance, one might know that a *samovar* is some kind of utensil associated with hot drinks, but be unsure about whether it is used for making the drink or for serving it. So one might be equally willing to accept that *samovar* signifies something like a kettle or an urn. By exposing subjects to paragraphs containing exemplars of *samovar* together with contextual cues lexically associated with each of these possible interpretations (i.e., *urn* vs. *kettle*), subjects' representations of the meaning of *samovar* may be nudged towards the meaning of the word associated with the contextual cues. Thus the dependent variable we would like to measure is the similarity of the two words' semantic representations.

While such a measurement is not directly possible, psychologists have developed a number of indirect methods that purport to tap into the semantic representations of words. We needed a task that would allow similarity in meaning to be reliably measured, while at the same time remain sensitive to the hypothesised changes in semantic representations due to the context manipulation. Similarity ratings meet these criteria, having a long history of use in psychological investigations of word meaning (e.g., Osgoode, Suci & Tannenbaum, 1957), and importantly, similarity judgements have been shown to be affected by context. For instance, Barsalou (1982) demonstrated that in a 'pets' context, the concepts *snake* and *raccoon* were judged to be more similar than if no context was provided. Medin, Goldstone and Gentner (1993) also observed context-dependent similarity effects: *black* was rated as more similar to *white* when also compared to *red* than when *black*⇔*white* was the only comparison required. We expected that subjects' ratings of between-word similarity, such as *samovar*⇔*kettle*, would be

Context A: 'urn'

On his recent holiday in Ghazistan, Joe slipped easily into the customs of the locals. In the hotel restaurant there was a samovar dispensing tea at every table. Guests simply served themselves from the samovar whenever they liked. Joe's table had an elaborately crafted samovar. It was the first earthenware samovar that he had seen.

Context B: 'kettle'

On his recent holiday in Ghazistan, Joe slipped easily into the customs of the locals. His hotel room featured a samovar and a single hob. Each morning Joe boiled water in the samovar for tea. Like others he had seen on his holiday, Joe's samovar was blackened from years of use. He imagined that at some point it would be replaced with an electric samovar.

Figure 1. The *urn*-biased and *kettle*-biased paragraph contexts created for *samovar*.

similarly influenced by the properties of the paragraph context which they had just read.

Method

Participants Forty-eight subjects, mostly undergraduate Psychology students at the University of Edinburgh, were recruited. All participants were native speakers of British English.

Materials and Design A list of 20 marginally familiar words (ten nouns and ten verbs) was compiled. Sixteen items were selected from the pre-tested materials used by Chaffin (1997) in his study examining free associations made to high- and low-familiarity words, and the remaining four were chosen by the authors. Items ranged in frequency from 0.13 to 2.92 occurrences per million (median: 0.64), according to a lemma frequency list created from the 100 million word British National Corpus (BNC).

For each item, we generated two 'target meanings' which we felt were plausible interpretations of the items. Then, for each of these target meanings we composed a short paragraph containing exactly four exemplars of the item. (See Figure 1 for a representative item with its paragraph contexts). Text passages were homogenous in structure, with the first sentence setting the scene; the marginally familiar words were embedded in the following three or four sentences. Passages ranged in length from 50 to 96 words (median length of 62). We attempted to bias the interpretation of the item in the paragraph by seeding the immediate context of each exemplar with strong lexical associates of the selected target meaning. For example, the meaning of *samovar* in Context B is 'pushed' towards *kettle* through the words *boiled*, *blackened* and *electric*, which are all more indicative of kettles than urns.

The strong lexical associates were generated in turn using a statistical technique commonly employed in computational linguistics for discovering collocations (e.g., Church & Hanks, 1990; Manning & Schütze, 1999); this procedure involved, for each target meaning (e.g., *urn*, *kettle*), collecting the co-occurrence frequencies of all words found in a ±5 word window around it in the BNC, converting these counts using the log-transformed odds ratio statistic (Agresti, 1990), and then sorting the resulting list. Strong associates – roughly, words that co-occur more often than expected by chance – tend to appear at the top of the ranking. We then selected suitable words for use as contextual cues from the topmost part of the list.

Paragraph contexts were randomly assigned to one of the two levels of the Context factor (A, B). This design is now sufficient to test for an effect of Context when subjects are asked to rate the similarity between e.g., *samovar* and *urn* after reading either Context A or Context B. In order to complete a factorial design, Context was crossed with a second factor, Target Meaning, with the same two levels, varying the word to which the marginally familiar item is compared.

The materials were next divided into four versions of 20 paragraphs each. Counterbalancing ensured that no participant saw the same item more than once.

Procedure Subjects were divided randomly amongst each of the four versions. The experiment was administered in the form of a questionnaire, with one paragraph context per page. Located below each paragraph was a numbered seven-point scale, and subjects were instructed to rate how similar the item was to the target meaning, where 'a 1 means "not at all similar" and a 7 means "highly similar"'; e.g., "How similar is a *samovar* to an *urn*?". The verb items were presented in present participle form; e.g., "How similar is *absconding* to *escaping?*". Order of presentation of the 20 items was randomised individually for each participant.

After completing the 20 items, subjects were required to rate a list of 28 words for familiarity, also using a 7-point scale, where 'a 1 means "very unfamiliar" and a 7 means "very familiar"'. This list comprised the 20 designated items plus eight filler words of moderate to high familiarity. The purpose of the familiarity ratings task was to allow a more detailed examination of the similarity data, in order to take into consideration the inherent variability in individuals' experience with the items.

Results

We conducted two-way repeated measures analyses of variance (ANOVAs) on the similarity judgements, treating both subjects and items as random factors.

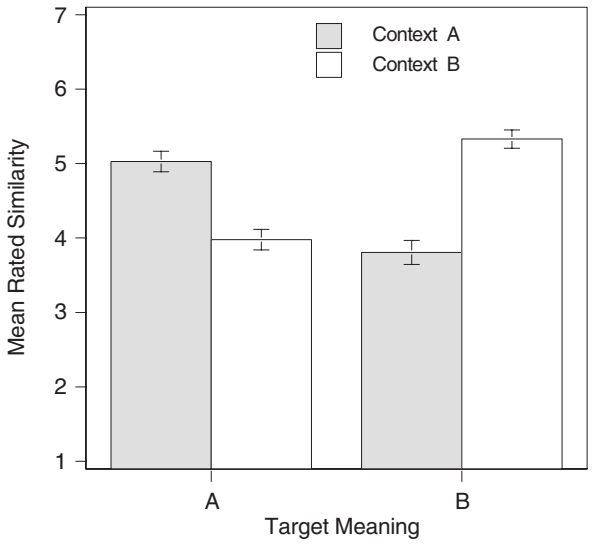


Figure 2. Mean semantic similarity as a function of Context and Target Meaning in Experiment 1.

There were no reliable main effects of either Target Meaning, $F_1(1,47)=4.02$, $MSE=0.667$, $p>0.05$; $F_2(1,19)=1.59$, $MSE=0.694$, $p>0.2$, or Context, $F_1(1,47)<1$; $F_2(1,19)<1$. The lack of a Target Meaning main effect indicates that, collapsing over the paragraph contexts in which the marginally familiar items were embedded, there was no bias between the 'A' and 'B' meanings in terms of their rated similarity to the item. The lack of a main effect of Context indicates an analogous absence of bias for the paragraph contexts.

There was a highly significant Context × Target Meaning interaction: $F_1(1,47)=60.04$, $MSE=1.323$, $p<0.001$; $F_2(1,19)=35.73$, $MSE=0.924$, $p<0.001$). As indicated by Figure 2, the interaction was due to Context effects at each level of Target Meaning. The mean similarity rating between a marginally familiar word and its 'A' meaning was higher when the item was embedded in the context biasing that meaning than when it appeared in the passage biasing the 'B' meaning.

Discussion

These results indicate that the distributional information contained in the paragraph contexts are sufficient to influence participants' similarity judgements. In the terminology of semantic space models, vectors were successfully 'pushed' towards other vectors in the representational space. Thus a strong prediction of the semantic space theory of meaning representation is supported: by selecting appropriate contextual cues and positioning them in the immediate linguistic context of a marginally familiar word, behavioural measures assumed to tap the word's meaningful properties can be influenced.

The results also provide support for the distributional hypothesis. Adding instances of a word in its environment of use to one's language experience – even as few as four exemplars – appears to be adequate to affect one's perception of its similarity in meaning to other words.

Although the items were chosen to be on the frontiers of familiarity for the subject population, the familiarity of a particular word can vary substantially between participants. For example, *samovar* may be a familiar word to someone who has travelled in Russia. According to the distributional hypothesis, this individual should be less influenced by the context when rating the similarity of *samovar* to *kettle* or to *urn*.

As we had collected familiarity ratings for each of the targets from each subject, we were able to address this question by dividing the ratings data points into low-familiarity (LoFam) and high-familiarity (HiFam) groups around the median familiarity score. The LoFam partition included data points with a self-rated familiarity score of three or less, and the HiFam group contained data for items rated as five or more.

The critical Context × Target Meaning interaction was present in the LoFam partition: $F_1(1,29)=59.24$, $MSE=1.80$, $p<0.001$; $F_2(1,17)=21.61$, $MSE=1.82$, $p<0.001$. The HiFam partition also displayed the interaction: $F_1(1,36)=21.55$, $MSE=1.80$, $p<0.001$; $F_2(1,17)= 30.28$, $MSE=0.92$, $p<0.001$.

It seems, then, that subjects' interpretations of marginally familiar words could be guided by the distributional properties of the contexts in which they were encountered, at least to the extent necessary to influence an immediately executed similarity rating. This effect was observed both for words with which subjects considered themselves reasonably familiar and for less familiar words.

The results of Experiment 1 raise two interesting questions with regard to our subjects' mental representations of the meanings of the stimuli: Were subjects actively using the distributional information in the contexts to actively augment (or even construct) their representation of the meaning of *samovar*? Or were the paragraph contexts activating particular features of their existing knowledge about samovars, causing the attendant shift in similarity ratings? In the latter case it could be argued that subjects' sensitivity to the distributional properties of words demonstrated in Experiment 1 is merely an epiphenomenon, a reflection of the fact that certain concepts share certain semantic features. On this account, the distributional properties associated with words arise *because* the concepts underlying the words possess certain features, and it is sensitivity to similarities between these concepts that subjects are actually manifesting. To examine these competing explanations, Experiment 2 controlled for the influence of any such prior conceptual knowledge by replacing Experiment 1's items with nonce words. Subjects were essentially starting from a 'tabula rasa' with respect to the meaning of nonce words, so evidence that the context was truly exerting an independent influence on subjects' judgements in Experiment 1 would be provided if similar effects of context are observed using nonce words.

Experiment 2

Experiment 2 controlled for the potential influence of participants' existing conceptual knowledge about the meaning of the target items by replacing the marginally

familiar items used in Experiment 1 with nonce words. (Thus the task now closely resembles the situation where an unknown word is encountered during reading, and its meaning has to be inferred from the context.)

Method

Participants Twenty subjects from the same population as Experiment 1 volunteered to take part.

Materials and Design The materials were identical to those used in Experiment 1, with the exception that the 20 marginally familiar items were replaced with orthographically-legal and pronounceable nonwords. For instance, all occurrences of *samovar* in the text passages were replaced with the nonce word *balak*. Care was taken that each nonce replacement did not phonologically resemble the original item or its two associated 'target meanings'.

Procedure The procedure was the same as for Experiment 1, except there was no familiarity ratings task.

Results and Discussion

Similarity ratings data were submitted to repeated measures ANOVAs. The Target Meaning × Context interaction was significant both by subjects: $F_1(1,19)=159.83$, $MSE=0.469$, $p<0.001$; and by items: $F_2(1,19)=40.23$, $MSE=1.863$, $p<0.001$. There were no main effects of either Target Meaning: $F_1(1,19)=1.09$, $MSE=0.385$, $p>0.3$; $F_2(1,19)<1$ or Context: $F_1(1,19)<1$; $F_2(1,19)<1$.

Thus these results are consistent with the findings of Experiment 1. It appears that any objections regarding the possible role and influence of prior knowledge about the meanings of Experiment 1's marginally familiar items are unfounded. Similarity comparisons involving unknown (nonce) words were also susceptible to manipulation of the same contextual cues that gave rise to the interaction in Experiment 1.

Simulating the Accumulation of Contextual Experience

Experiments 1 and 2 have shown that a very small amount of experience with a word in context is capable of influencing similarity judgements involving that word. The items in Experiment 1 were selected to represent the sorts of words to which subjects would be expected to have a low level of prior exposure. If it were possible to *increase* the amount of one's prior contextual experience with a given item, the influence of subsequent exposure (i.e., the four-exemplar paragraphs in Experiment 1) should be reduced. We simulated this effect of previous experience using a semantic space model derived from distributional statistics. We predicted that the size of the simulated context effect would diminish as the ratio of previous experience to the experience provided by the paragraphs increased. We varied the amount of contextual exposure given to the model by varying the size of the corpus used to construct co-occurrence vector representations for the 20 marginally familiar items.

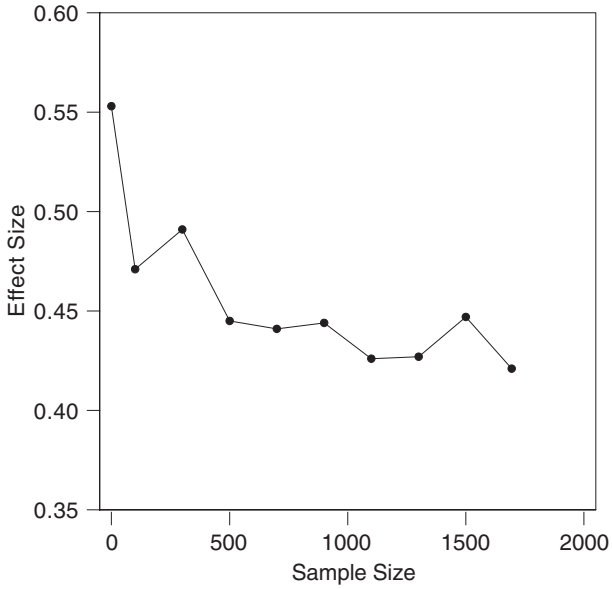


Figure 3. The size of the Consistency effect as a function of the amount of contextual experience.

Method

From the BNC, we extracted the ±5 word contexts surrounding every occurrence of all 20 items (a total of 1,694). We then took random samples (with replacement) of various sizes from this item-in-context 'corpus', appending them to both an analogous corpus formed by the 'A' passages and the corpus formed by the 'B' paragraphs, resulting in separate 'A' and 'B' corpora for each sample size.

From each 'corpus', we extracted co-occurrence vectors for the 20 items using a window size of ±5 words and the 20,000 most frequent content words as context words. The resulting item vectors thus directly reflect the ratio of previous experience to subsequent experience (vectors created from the passages only simulate a complete lack of previous experience with the word). Vectors for the 40 'target meanings' (e.g., *urn*, *kettle*) were constructed using the entire BNC.

Results and Discussion

We collapsed the 2 × 2 design of Experiment 1 into a single factor, Consistency, in order to compare the vector similarity of an item with each of its 'target meanings', between the case where the paragraph context is consistent with (or biases) the target meaning (e.g., *samovar*⇔*urn* for Context 'A'; see Figure 1) and the case where it is inconsistent (*samovar*⇔*urn* for Context 'B'). Similarity was computed as the cosine of the angle between vectors, and a paired-*t* test was conducted on the cosine measurements. Consistent comparisons should return a larger cosine than Inconsistent comparisons. At the $\alpha=0.01$ level of significance, reliable Consistency effects were observed for all sample sizes but one (the effect for the 1100-exemplar sample was significant at $\alpha=0.05$).

In order to illustrate the effect of increasing the amount of previous experience, Figure 3 displays the Consistency effect size (Cohen's *d*) as the sample size varies. As expected, the effect is largest for vectors created from the passages only, and diminishes as more contextual experience is added. Both Experiment 1's results and the anticipated effect of variable amounts of prior exposure were simulated in a semantic space model drawing only upon distributional information.

General Discussion

To summarise, manipulating the contextual cues present in short text passages was sufficient to influence adults' similarity judgements involving marginally familiar and nonce words embedded in these passages. Our results suggest that readers' interpretations of these items were 'pushed' towards the meanings of other words. Analogous to the way that the meaning of unknown words can be determined while reading, contextual information is also an influential factor when consolidating the meaning of words on the frontiers of familiarity.

The experimental results also suggest that a remarkably small amount of exposure to a word in a meaningful context is sufficient to influence similarity ratings. However, the relative recency of this experience is likely an important factor; the context effect may well diminish as a function of the length of time between reading the paragraph and making the similarity judgement.

Though a simple model of word learning, the semantic space simulation illustrated the decrease in susceptibility to contextual manipulation expected as one's prior experience with a word increases. Of course, we do not claim that human semantic space has 20,000 dimensions; rather, what is important is the inferences that can be drawn about a word's meaning simply by taking note of the words in its immediate context. It is notable that the simulated Consistency effect was still reliable even after all the contextual experience in the BNC was added; in as much as the BNC can be considered to represent the average person's language exposure, it seems that very little extra contextual experience is needed to affect the perception of a word's similarity in meaning to other words.

Acknowledgements

We thank Dan Yarlett for useful discussion and comments.

References

Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley & Sons.

Barsalou, L. W. (1982). Context-independent and context-dependent information in concepts. *Memory & Cognition*, *10*, 82-93.

Carnine, D., Kameenui, E. J. & Coyle, G. (1984). Utilisation of contextual information in determining the meaning of unfamiliar words. *Reading Research Quarterly*, *19*, 188-204.

Chaffin, R. (1997). Associations to unfamiliar words: Learning the meanings of new words. *Memory & Cognition*, *25*, 203-226.

Church, K. W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*, 22-29.

Fischer, U. (1994). Learning words from context and dictionaries: an experimental approach. *Applied Psycholinguistics*, *15*, 551-574.

Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. Boston: Kluwer.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211-240.

Lowe, W. & McDonald, S. (2000). The direct route: Mediated priming in semantic space. *Proceedings of the 22th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Lund, K., Burgess, C. & Atchley, R. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 660-665). Mahwah, NJ: Erlbaum.

Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press

McDonald, S. (2000). *Environmental determinants of lexical processing effort*. Doctoral dissertation, Division of Informatics, University of Edinburgh.

McDonald, S. & Lowe, W. (1998). Modelling functional priming and the associative boost. *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 667-680). Mahwah, NJ: Erlbaum.

Medin, D. L., Goldstone, R. L. & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*, 254-278.

Osgoode, C. E., Suci, G. J. & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Patel, M., Bullinaria, J. A. & Levy, J. P. (1998). Extracting semantic representations from large text corpora. In J. A. Bullinaria, D. Glasspool & G. Houghton (Eds.) *Proceedings of the 4th Neural Computation and Psychology Workshop, London, 9-11 April 1997*. London: Springer-Verlag.

Ramscar, M. J. A. & Yarlett, D. G. (2000). The use of a high-dimensional, "environmental" context space to model retrieval in analogy and similarity-based transfer. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 381-386). Mahwah, NJ: Erlbaum

Redington, M., Chater, N. & Finch, S. (1998). Distributional information: a powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*, 425-469.

Saffran, J. R., Aslin, R. N. & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926-1928.

Saffran, J. R., Newport, E. L. & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606-621.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, *24*, 97-124.

Activating Verbs from Typical Agents, Patients, Instruments, and Locations via Event Schemas

Ken McRae (kenm@uwo.ca)
Department of Psychology, Social Science Centre
University of Western Ontario, London, ON, Canada N6A 5C2

Mary Hare (hare@rowan.bgsu.edu)
Department of Psychology, Bowling Green State University
Bowling Green, OH 43403 USA

Todd R. Ferretti (tferrett@cogsci.ucsd.edu)
Jeffrey L. Elman (elman@cogsci.ucsd.edu)
Department of Cognitive Science, UC San Diego
9500 Gilman Drive, La Jolla, CA 92093 USA

Abstract

Verbs are a major component of theories of language processing, partly because they exhibit systematic restrictions on their arguments. However, verbs follow their arguments in many constructions (particularly in verb-final languages), making it inefficient to defer processing until the verb. Computational modeling suggests that during sentence processing, nouns may activate information about subsequent lexical items, including verbs. We investigated this prediction using short stimulus onset asynchrony (SOA, the time between the onsets of the prime and target) priming. Robust priming obtained when verbs were named aloud following typical agents (nun - praying), patients (dice - rolled), instruments (shovel - digging), and locations (arena - skating). This research is the first to investigate systematically the priming of verbs from nouns. It suggests that event memory is organized so that entities and objects activate the class of events in which they typically play a role (Lancaster & Barsalou, 1997). These computations may be an important component of expectancy generation in sentence processing.

Introduction

Expectancy generation is a central construct in many theories of language comprehension, although the term has been used in a variety of ways. In the current work, we use it to refer not to explicit generation of expectancies (as in Becker's, 1980, verification model) but rather to the implicit expectancy generation that is a natural by-product of language comprehension processes (as exemplified, for example, by simple recurrent networks and other types of networks that implement processing through time, or by incremental parsing models with a predictive component). Consistent with the notion that comprehenders implicitly generate expectations, a number of computational models and human experiments have shown that both local and global context can narrow expectations for upcoming words in a sentence (e.g. Elman, 1990; Schwanenflugel & Shoben, 1985).

Expectancy generation in theories of sentence processing has focused primarily on verbs. This is reasonable because verbs, as predicating functions, tend to exhibit regular and systematic restrictions on the arguments with which they co-occur. In line with this notion, researchers have concentrated on the ways in which people's knowledge of verb argument structure and of the thematic roles associated with verbs constrains what may follow. It has been shown, for example, that a verb can restrict the range of syntactic structures or broad classes of words that are likely to follow it. One clear example of this is the influence of verb subcategorization preferences on resolving the direct-object/sentence complement ambiguity (Garnsey et al., 1997).

Thematic Roles

Verbs and their associated thematic roles are a major component of most linguistic and psycholinguistic theories of language processing, and many psycholinguistic experiments have focused on thematic role assignment. One reason for this is that thematic roles are hypothesized to be an important locus of semantic/syntactic interaction (Tanenhaus & Carlson, 1989). Recently, McRae, Ferretti, and Amyote (1997) incorporated and extended the important work of Carlson and Tanenhaus (1988), Dowty (1991), and others to construct a theory of thematic roles that incorporates event-specific information. On this account, thematic roles are viewed as including verb-specific concepts, and this conceptual/world knowledge is computed and used immediately in on-line language

processing (see Altmann & Kamide, 1999, for supportive evidence from head-mounted eyetracking). As part of this research, Ferretti, McRae, and Hatherell (2001; Experiments 1 and 2) showed that verbs in isolation prime typical agents, patients, and instruments. Their Experiment 4 showed further that priming was limited to the appropriate role when active versus passive sentence fragments were used to cue roles. Given these conjoint effects of semantic and syntactic cues, they concluded that verbs activate event schemas and that this knowledge should be considered as part of thematic roles.

Expectancies from Nouns to Verbs

Because of the emphasis on verbs in the sentence processing literature, one often finds characterizations in which processing is dependent to a high degree on the verb, implying that processing may be held in abeyance until the verb is heard or read. However, in verb-final languages such as German, deferring any hypothesis about structure and meaning until the end of the sentence or clause would be an inefficient processing strategy (Frazier, 1987). Furthermore, a number of recent articles have addressed incremental syntactic processing in verb-final constructions in German and Japanese (e.g., Kamide & Mitchell, 1999).

There are reasons to believe that under many circumstances, words from major syntactic categories other than verbs may exert powerful constraining forces on expectancy generation. For example, nouns can possess valence restrictions, although they do so to a lesser degree than do verbs. In addition, our work with modeling suggests that network dynamics will encode the degree to which elements other than the verb restrict the range of what may follow. Therefore, non-verbs may generate powerful expectations about what might follow, including expectations about possible verbs.

Strong candidates for driving expectations for semantic classes of verbs are typical fillers of their thematic roles: that is, the agents, patients, and instruments that often are involved in specific types of events, as well as locations at which specific events typically occur. The present research was designed to test this possibility. The logic underlying the experiment is that certain nouns may produce expectations for certain semantic classes of verbs during normal on-line sentence comprehension by activating event schema knowledge. If nouns do activate event knowledge associated with verbs, then those nouns should prime corresponding verbs in a short SOA single-word priming task. Note that we are not claiming that expectancy generation necessarily drives performance in a short SOA priming task. Instead, the claim is that nouns activate event schemas, and this drives the priming. The implication is that this knowledge, once activated, can then serve as a source for expectancy generation during on-line comprehension of full sentences in natural language.

Before describing the experiment itself, we briefly outline another aspect of the theoretical motivation, based on recent work on event memory.

Autobiographical Event Memory

In much the same way that language researchers have focused on knowledge computed from verbs, theories of autobiographical memory often have focused on activity as a primary organizing principle of event representations. Some researchers have characterized this as the strong activity view, whereby events are organized and accessed by activity only. Thus in parallel with the emphasis on verbs in the psycholinguistic literature, much of the literature on the organization of autobiographical memory emphasizes the centrality of events, which often are realized linguistically as verbs. The Ferretti et al. (2001) results showing that some verbs provide strong expectations for classes of noun concepts to fill specific thematic roles could be viewed as on-line support for this position.

At the same time, the literature on the organization of event memory also suggests that nouns may produce expectations for specific classes of verbs. Lancaster and Barsalou (1997) found that people are adept at organizing short narratives in terms of multiple components of events, including activity (i.e., by verb), time, participants (i.e., agents and patients), and location. They argued that contrary to the strong activity view, memory is organized to allow access from multiple components of events.

Our argument assumes that people's knowledge of a generalized event such as skating is constructed over time from individual event instances, and can be computed in multiple ways. If this is correct, nouns should quickly activate well-learned event knowledge, so that typical agents, patients, instruments and locations should prime verbs denoting the event.

Experiment

The present research tested this prediction using short SOA priming from nouns to verbs. The noun primes referred to entities, objects, and locations that are typically involved in the events denoted by the target verbs. We predicted shorter naming latencies for verbs primed by their common thematic role fillers than for verbs primed by unrelated nouns because a related noun will activate the schema corresponding to the type of event in which it typically participates, thus activating a verb denoting that type of event.

Method

Participants. Forty University of Western Ontario undergraduates participated for course credit. All participants were native speakers of English and had normal or corrected-to-normal visual acuity.

Materials. To tap into people's knowledge of the types of events in which certain entities and objects play a specific role, we used what we will refer to as thematic-based event generation norms. These norms are designed to estimate the conditional probability of a generalized event given an entity or object playing a specific role. Participants were asked to generate verbs in response to typical agents, patients, instruments, and locations. In the agent norms, participants were given nouns such as nun and were asked to "List the things that these people commonly do." In the patient norms, participants were given nouns such as dice and asked to "List the things that these objects/people commonly have done to them." In the instrument norms, participants were given nouns such as shovel and asked to "List the things that people commonly use each of the following to do." Finally, in the location norms, participants were given nouns such as arena and asked to "List the things that people commonly do at/in each of these locations." both For each item, ten blank lines were provided for responses. No time limit was imposed. Participants were undergraduate students from Bowling Green State University. Each participant completed only one list; there were approximately 25 items per list. In total, 20 participants responded to each item.

Responses were scored based on their rank order within a participant, and on their response frequency. That is, each response was scored in terms of the number of participants listing it first, second, third, through to tenth. A weighted score was calculated for each response by multiplying the frequency with which it was produced first times 10, second times 9, and so on, and then summing those products. Wherever possible, noun-verb pairs were chosen for the priming experiment by taking the verb with the highest weighted score. In a few cases, the response with the highest weighted score could not be used because it was a multi-word phrase, such as work out for gymnasium, and the constraints of the naming task demanded a single-word verb target. In a few other cases, the same verb was the best response for more than one item (e.g., cut for both chainsaw and knife). In both of these cases, the verb chosen for the priming experiment was either the next-best response, or a synonym or near synonym of the best response. For example, because cutting was used as the target for chainsaw, slicing was used instead of cutting as the target for knife.

From these norms, we chose 30 agents paired with the present participle of a verb such as nun - praying, waiter - serving, and lawyer - defending, 30 patients paired with the past participle form of a verb, such as teeth - brushed, dice - rolled, and tax - paid, 32 instrument-present participle pairs such as shovel - digging, pen - writing, and chainsaw - cutting, and 24 location-present participle pairs such as cafeteria - eating, bedroom - sleeping, and bathroom - showering. The weighted scores for the verbs for each thematic role were (maximum of 200): agents, $M = 91$, $SE = 10$; patients, $M = 111$, $SE = 10$; instruments, $M = 134$, $SE = 8$; and locations, $M = 114$, $SE = 11$.

Target verbs were presented in present participle form with the agent, instrument, and location primes. Verb targets paired with typical patients, however, were presented in their past participle forms to avoid including prime-target pairs that formed coherent familiar phrases, such as cigar smoking.

Lists. Two lists were created for each type of noun (i.e., agents, patients, instruments, and locations). Each list contained half of the related and the opposing half of the unrelated items. Unrelated items were created by re-pairing the nouns and verbs from the related trials in the opposite list. Filler trials consisted of unrelated noun-verb pairs such as stapler vacuuming. Each list contained four times as many unrelated filler trials as related target items (relatedness proportion was .17). Thirty-five unrelated practice trials were used for the practice session for every participant. No participant saw any word twice.

Procedure. For each trial, the participant was instructed to read silently the first word presented on the computer screen and to pronounce aloud the second word as quickly and accurately as possible into the microphone. Stimuli were presented on a 14-inch Sony Trinitron monitor connected to a Macintosh LC630 using PsyScope (Cohen et al., 1993). A microphone connected to a CMU button box measured naming latency (in ms) as the time between the onset of the target and the participant's pronunciation of it. Each trial consisted of: a focal point (*) for 250 ms; the prime for 200 ms; a mask (&&&&&&&&&) for 50 ms; and the target until the participant named it. The intertrial interval was 1500 ms, and a break was given every 40 trials. Testing sessions began with the practice trials and lasted approximately 20 minutes. The experimenter recorded trials in which the participant mispronounced a word (a pronunciation error), extraneous noise caused the voice key to trigger (a machine error), or the voice key failed to trigger (a machine error). Participants were assigned randomly to be tested on either the agents and locations, or on the patients and instruments. The order of the two lists was counter balanced across participants (i.e., ten participants were tested on agents then locations, and another ten on locations then agents, and the same for the patients and instruments).

Design. Naming latencies and the square root of the number of pronunciation errors (Myers, 1979) were analyzed by separate two-way analyses of variance for each thematic role (agents, patients, instruments, and locations). The factor of interest was relatedness (related vs. unrelated), which was within both participants (F_1) and items (F_2). List was included as a

Table 1: Mean Verb Naming Latencies (ms) and Percentage Pronunciation Errors

	Agents		Patients		Instruments		Locations	
Dependent Measure	M	SE	M	SE	M	SE	M	SE
Response Latency								
Unrelated	592	21	583	20	565	20	578	16
Related	574	19	561	18	549	17	560	19
Facilitation	18*		22*		16*		18*	
Percentage Errors								
Unrelated	1.9	0.8	3.2	1.4	1.4	0.7	2.5	1.2
Related	1.9	0.9	1.9	0.8	1.1	0.6	1.5	0.8
Facilitation	0		1.3		0.3		1.0	

* Significant by participants and items

between-participants dummy variable and item rotation group as a between-items dummy variable to stabilize variance that may result from rotating participants and items over the two lists (Pollatsek & Well, 1995).

Results

Naming latencies greater than three standard deviations above or below the grand mean were replaced by that value (1% of trials). Two participants were dropped because their soft speaking style resulted in an extreme number of trials in which the voice key was not activated. Machine errors, the majority of which were caused by the microphone failing to register the participant's response, occurred on 4% of the trials, were excluded from all analyses. Pronunciation errors were excluded from the latency analyses. Mean naming latency and percent pronunciation errors are presented for each condition in Table 1. Verbs were named more quickly when preceded by a related versus an unrelated noun for each of the four thematic roles: agents: $F_1(1,18) = 6.19$, $p < .05$, $F_2(1,28) = 4.12$, $p < .06$; patients: $F_1(1,18) = 7.54$, $p < .05$, $F_2(1,28) = 11.98$, $p < .001$; instruments: $F_1(1,18) = 5.66$, $p < .05$, $F_2(1,30) = 7.64$, $p < .01$; and locations: $F_1(1,18) = 5.33$, $p < .05$, $F_2(1,22) = 10.41$, $p < .01$. There were no reliable differences in pronunciation error rates, all F's < 1.

Discussion

Noun-verb pairs were chosen using thematic-based event generation norms designed to tap into people's knowledge of the conditional probability of a generalized event given an agent, patient, instrument, or location. Significant noun-verb priming was found in all four cases. To our knowledge, this is the first experiment to investigate systematically the priming of verbs from nouns.

This experiment shows that nouns make available information about events in which they typically play a role. One plausible explanation of these results is that, as in the weak activity view of event memory (Lancaster & Barsalou, 1997), event schemas are organized so that they are accessible from common agents, patients, instruments and locations. That is, mental representations of generalized events are structured so that they can be computed quickly when a noun that refers to a typical component of a specific type of event is read or heard. When this generalized event knowledge is computed, the verb corresponding to this type of event is partially activated, thus resulting in the priming effects found in the Experiment. In other words, language and event memory are organized so that event knowledge can be accessed quickly from nouns, as well as from verbs. This explanation is consistent with Moss et al. (1995) who found priming using functionally-related items such as broom-floor. Moss et al. concluded that priming in their experiment occurred through representations of generalized events.

The fact that nouns can activate information about corresponding verbs suggests that, at least in some circumstances, nouns may be a strong source of expectancy generation for ensuing verbs. This may be particularly important for languages such as German and Japanese, which contain numerous verb-final constructions, but it may also play a key role in sentence comprehension (and production) in English. At least one noun phrase precedes the verb in the vast majority of English utterances, and in many constructions the verb appears late in the clause, as in questions (Which customer did he serve?) and it-clefts (It was vase on the coffee table that she broke.).

One question that might be asked of the present results is why priming was found from locations to verbs, whereas Ferretti et al. (2001) failed to find

priming from verbs to locations. The most likely explanation concerns the type of norming used in the two experiments. The present study used thematic-based event generation norms designed to tap people's knowledge of the conditional probability of an event given (in this case) a location. In contrast, Ferretti et al. used role/filler typicality norms in which participants were asked to provide ratings for questions such as "How common is it for someone to draw in each of the following locations?" This may not be the best way to measure the conditional probability of the location given the event. For example, although the mean role/filler typicality rating for draw-studio was 6.5 out of 7, there are numerous locations where drawing can occur, and a studio might not rank as the best.

To test this possibility, we conducted a further set of norms, parallel to the event generation norms used in the current study. Participants listed locations at which some event might occur, providing an estimate of the conditional probability. In these norms, the mean weighted score for Ferretti et al.'s (2001) verb-location items was only 44 (maximum of 200). In contrast, the mean weighted score of the location-verb items used in the present research was 114. This difference between the conditional probabilities in the relevant directions may account for the discrepancy between studies.

Note, however, that the Ferretti et al. (2001) study did find robust priming in the other three verb-noun conditions (i.e., verbs primed agents, patients, and instruments), raising the possibility that our results may be due to backward priming from the verb. If this was the case, it would seriously reduce the theoretical import of our results. In response to this potential criticism, we note that Kahan, Neely, and Forsythe (1999) and Peterson and Simpson (1989) have shown that backward priming occurs in a naming task at a short SOA, but not at an SOA of 500 ms. Therefore, we currently are replicating this experiment using a 500 ms SOA. If verbs are primed, which we expect they will be given the conditional probabilities as evidenced by our norms, then we can be positive that backward priming is not responsible for our results.

Spreading Activation Networks. A common explanation of word-word priming results focuses on spreading activation in a semantic network. Therefore, an interesting question concerns whether spreading activation networks would predict priming of verbs from agents, patients, instruments, and locations. Although the first semantic networks focused solely on noun representations (Collins & Quillian, 1969), relatively early extensions incorporated verbs (Gentner, 1975; Rumelhart & Levin, 1975). Verb representation included core meaning plus thematic links to nodes that stood as placeholders for possible noun phrases that fill those roles in sentences. Note that these links could be bi-directional, so that activation could spread from the thematic role nodes to the verb node. However, these thematic links and nodes included minimal semantic content that was restricted to general selectional restriction information. For example, a thematic link between a verb and an agent node might specify that the filler of that node must be animate. Thus, spreading activation models of this type predict no priming because the experiment reported herein controlled for general selectional restrictions in that the related and unrelated trials were equivalent in terms of this factor.

If current semantic networks were expanded, it could be assumed that noun nodes representing common agents, patients, instruments, and locations become linked to specific verb nodes over time. These links might be formed because of people's experience with events (via noticing that chainsaws are used for cutting), and/or linguistic descriptions of events (e.g., via word co-occurrence in speech and text). In this view, when participants read the noun prime in our experiment, activation might spread to the verb node and priming might result. Thus spreading activation networks could predict priming of verbs from typical thematic role fillers, but only by incorporating ad hoc assumptions well outside the scope of current versions of the theory.

Undifferentiated links encoding associative relatedness provide a possible second way in which spreading activation networks might predict priming from typical thematic role fillers to verbs. If the representations of words and/or concepts that often co-occur in events and language become linked in one or more of semantics, orthography, and phonology via an unspecified associative relation, then those links could serve as the basis for priming from nouns referring to typical components of events to verbs. Theoretically, however, it is a step backward to treat this knowledge as an undifferentiated associative relation because it is known to be thematic-driven knowledge concerning the relationship between generalized events and their common components.

Finally, a recent experiment, part of this line of research, produced results that are extremely difficult for a spreading activation network to account for, even given the additional ad hoc assumptions described above. Ferretti (2000) manipulated verb aspect to reference various components of the temporal structure of events. In one condition, he presented a verb prime in its imperfective aspect (e.g., was skating), which references an event as ongoing. Ferretti reasoned that if an event is in the process of occurring, then the location at which the event is taking place should be salient. In the second condition, the verb prime was presented in its past perfect form (e.g., had skated), which references the event as completed. That is, perfect aspect focuses on the resultant states of an event. Ferretti reasoned that if the event is referred to as completed, the location at which that type of event typically occurs should not be as salient. Thus, if priming is due to event knowledge and aspect references various components of the

temporal structure of events, priming should be found with imperfective aspect (where location is salient) but not with perfect aspect (where it is not). In a spreading activation network account, there should be no influence of aspect. In a short SOA priming task, typical locations were significantly primed by verbs presented in their imperfective aspect, whereas no priming obtained when perfect aspect was used. These results follow naturally from an account in which priming occurs via event schemas. However, accounting for them in a spreading activation network requires incorporating some mechanism by which aspect can modulate the flow of activation.

Conclusions

Event memory is organized in multiple ways, making it accessible from nouns as well as from verbs. Because of this, agents, patients, instruments, and locations prime verbs denoting events in which they typically play a specific role. These results suggest that nouns can be a basis for generating expectancies for upcoming verbs during on-line sentence comprehension.

Acknowledgments

This research was supported by NSERC (Canada) grant OGP0155704 and a PREA (Ontario) grant to the first author, and NSF grant DBSS92-09432 to the fourth author.

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. Cognition, 73, 247-264.

Becker, C. A. (1980). Semantic context effects in visual word recognition: An analysis of semantic strategies. Memory & Cognition, 8, 493-512.

Carlson, G. N., & Tanenhaus, M. K. (1988). Thematic roles and language comprehension. In W. Wilkins (Ed.), Thematic relations (pp. 263-288). New York: Academic Press.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. Behavioral Research Methods, Instruments & Computers, 25, 257-271.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior, 8, 240-247.

Dowty, D. (1991). Thematic proto-roles and argument selection. Language, 67, 547-619.

Elman, J. L. (1990). Finding structure in time. Cognitive Science, 14, 179-211.

Ferretti, T. R. (2000). Situation schemas, thematic roles and grammatical morphemes. Unpublished Doctoral Dissertation. The University of Western Ontario.

Ferretti, T. R., McRae, K., & Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. Journal of Memory and Language, 44, 516-547.

Frazier, L. (1987). Syntactic Processing: Evidence from Dutch. Natural Language and Linguistic Theory, 5, 519-560.

Garnsey, S. M., Pearlmutter, N. J., Meyers, E., & Lotocky, M. A. (1997). The contribution of verb-bias and plausibility to the comprehension of temporarily ambiguous sentences. Journal of Memory and Language, 37, 58-93.

Gentner, D. (1975). Evidence for the psychological reality of semantic components: The verbs of possession. In D. A. Norman and D. E. Rumelhart (Eds.), Explorations in Cognition (pp. 211-246). San Francisco: W. H. Freeman.

Kahan, T. A., Neely, J. H., & Forsythe, W. J. (1999). Dissociated backward priming effects in lexical decision and pronunciation tasks. Psychonomic Bulletin & Review, 6, 105-110.

Kamide, Y., & Mitchell, D. C. (1999). Incremental pre-head attachment in Japanese parsing. Language and Cognitive Processes, 14, 631-662.

Lancaster, J. S., & Barsalou, L. W. (1997). Multiple organizations of events in memory. Memory, 5, 569-599.

McRae, K., Ferretti, T. R., & Amyote, L. (1997). Thematic roles as verb-specific concepts. Language and Cognitive Processes, 12, 137-176.

Moss, H. E., Ostrin, R. K., Tyler, L. K., & Marslen-Wilson, W. D. (1995). Accessing different types of lexical semantic information: Evidence from priming. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 863-883.

Myers, J. L. (1979). Fundamentals of experimental design. Boston, MA: Allyn & Bacon.

Peterson, R. R., & Simpson, G. B. (1989). Effect of backward priming on word recognition in single-word and sentence contexts. Journal of Experimental Psychology: Learning, Memory, & Cognition, 15, 1020-1032.

Pollatsek, A., & Well, A. D. (1995). On the use of counterbalanced designs in cognitive research: A suggestion for a better and more powerful analysis. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 785-794.

Rumelhart, D. E., & Levin, J. A. (1975). A language comprehension system. In D. A. Norman and D. E. Rumelhart (Eds.), Explorations in Cognition (pp. 179-208). San Francisco: W. H. Freeman.

Schwanenflugel, P., & Shoben, E. J. (1985). The influence of sentence constraint on the scope of facilitation for upcoming words. Journal of Memory and Language, 24, 232-252.

Tanenhaus, M. K., & Carlson, G. (1989). Lexical structure and language comprehension. In W. D. Marslen-Wilson (Ed.), Lexical Representation and Process (pp. 505-528). Cambridge, MA: MIT Press.

Spatial Experience, Sensory Qualities, and the Visual Field

Douglas B. Meehan (dbmeehan@yahoo.com)
Department of Philosophy, CUNY Graduate Center
365 Fifth Avenue
New York, NY 10016-4309 USA

Abstract

An explanation of the qualitative nature of visual experience must account for the spatial character of visual sensations. We need mental locations and a mental visual field to explain how visual sensations enable us to perceive physical entities as located, to explain the differences in the spatial characters of sensations revealed in introspection, and to explain the spatial character of visual anomalies such as nonillusory afterimages. But mental locations are properties of mental states, not objects, so they must not be spatial locations. I argue that the homomorphism view of sensory qualities, pioneered by Wilfrid Sellars (1956, 1959, 1960, 1967) can best explain the nature of mental space. This view explains mental locations as nonspatial mental counterparts of physical locations. Austen Clark (1996, 2000) rejects mental location, claiming that the spatial character of sensory experience can be explained in terms of the spatial properties of distal stimuli and of sensory receptors. I argue that Clark's theory fails in two respects: 1) it fails to explain how it is that we individuate our sensory experiences with respect to apparent spatial differences, and 2) we cannot make the psychophysical identifications needed to get Clark's theory off the ground without first picking out the sensations by their mental properties.

Introduction

The current philosophical literature on visual experience is rife with debate on the issue of the qualitative character of sensory experience. The debate focuses largely on the topic of color and color qualia. But an explanation of color qualia will not exhaust a philosophical account of the qualitative nature of visual experience. We must also explain the nature of the spatial character of visual experience. Just as there is a qualitative difference between seeing something red and seeing something green, there is a qualitative difference between seeing a red patch off to one's right and seeing a red patch off to one's left. What it is like to see a red patch off to one's right is different from what it is like to see a red patch off to one's left.

But what are these qualitative spatial properties and relations? Are they actually spatial? When one has a visual experience of a red patch to the left of another red patch are there two phenomenal red entities located next to one another? And if they are located in any sense, where are they located? Are they located in some mental space—a visual field?

I argue that we do need to posit a mental space or visual field to explain the spatial character of visual experience. And I argue that the homomorphism view of sensory qualities, pioneered by Wilfrid Sellars (1956, 1959, 1960, 1967)[1] offers the best solution. In so doing I examine and reject Austen Clark's (1996, 2000) recent repudiation of mental space.

1. From Sensations to Sensory Fields

At the most basic level, we become conscious of our surroundings by having sensations that represent those surroundings. Sensations have perceptual roles. We see a red firetruck because we have a sensation of a certain type. We see a yellow Volkswagen because we have a sensation of another type. We see a red firetruck to the left of a yellow Volkswagen because we have a sensation of yet another type. Visual sensations enable us to perceive the firetruck next to the Volkswagen.

If we detect differences in visual stimuli by having different sensations, those sensations must vary in ways corresponding to perceptible differences in visual stimuli. Since we see entities by seeing differences in color, visual sensations must have properties corresponding to physical color properties. I will call these properties mental colors. I adopt Sellars's notational device of suffixing a '*' to color predicates to indicate reference to mental colors—e.g., red* is the sensory quality that enables us to perceive surfaces we call red.

We also perceive visual stimuli as having spatial properties. That we perceive the red firetruck as being to the left of the yellow Volkswagen seems to entail that our visual sensations have properties corresponding to the locations of the firetruck and Volkswagen. If sensations enable us to perceive stimuli as being located, sensations must have properties corresponding to locations. They must have mental locations, or locations*. Seeing an entity off to the right requires that one has a visual sensation with a property corresponding to being off to the right. Thus, we can infer that there are mental locations from the fact that

[1] The homomorphism view has also been argued for more recently by David Rosenthal (1998, 1999, 2000) and Sydney Shoemaker (1975). The version I argue for is closest to Rosenthal's.

we can perceive distal stimuli as occupying physical locations.

We can also introspect our sensations. When we do, we are conscious of them as having certain qualities. For instance, when introspecting the sensation that enables us to perceive the red firetruck to the left of the yellow Volkswagen, we are conscious of the sensation as being of a certain type—a sensation of a big red patch to the left of a little yellow patch. Introspection reveals the ways that sensations differ from one another, so it reveals something about the mental properties of those sensations. Inasmuch as sensations can differ with respect to apparent color and apparent location, they must have colors* and locations*.

In fact, we have sensations even in the absence of the appropriate stimuli. For instance, I can have a sensation of red at the center of my visual field without there being anything red directly in front of my eyes. We might claim that I am in the kind of sensation normally had when there is a red thing in the center of my distal visual field. Inasmuch as this state differs from the kind of sensation normally had when there is a red thing in the periphery of my distal visual field, we need to explain how it differs. This suggests some property of the sensations corresponding to the locations of their normal distal causes.

A strong example of sensations in the absence of stimuli is the phenomenon of nonillusory afterimages. A nonillusory afterimage is a sensation of a bright pattern one has after a flashbulb goes off in one's eyes. The afterimage seems to occlude objects in the distal visual field. When one moves one's eyes, the afterimage appears to occlude different objects.

But these afterimages are nonillusory. One does not think that there is some bright patterned object moving along with one's eyes. One thinks that it just appears that way. As Paul Boghossian and David Velleman put it, "The after-image must … be described as appearing in a location without appearing to be in that location ..." (1989, p. 91). An explanation of this fact, according to Boghossian and Velleman, is that the afterimage is located in a visual field. This visual field overlays the distal visual field, so the afterimage occludes whatever its region of the visual field happens to overlay. Since an afterimage appears in a location without appearing to be in a location, there must be mental space in addition to physical space.

But how do we explain mental space? Is it actually spatial? Is the afterimage located on some two dimensional transparency? This hardly seems feasible. If the mental visual field is an overlay through which we look at the world, we still need to explain how it is that we look through the overlay. Such an explanation must posit further states to mediate our seeing through the overlay. This will require another overlay, and so on ad infinitum.

Further, sensations are mental states, or events, not objects. Although a state is a state of some entity located somewhere, it is not clear how that location helps explain differences in the apparent locations of sensations.

If we must posit mental space, we must explain it as nonspatial. And if mental space is nonspatial, we must explain how mental locations correspond to spatial locations. The homomorphism view of sensory experience meets these challenges.

2. The Homomorphism View

The homomorphism view explains the relationship between a physical stimulus property and its mental counterpart property in terms of a common structure between their respective quality families. I will motivate the view with respect to color vision, and I will then explain how it extends to spatial experience.

According to the homomorphism view, physical red and red* are not the same property, nor do they resemble each other. They are counterparts in virtue of a similarity between the property families of which they are members—the color and color* families.

A property family is comprised of properties that resemble and differ from one another in varying degrees. For instance, the color family is comprised of perceptible colors.[2] Red is more similar to orange than it is to green. Green is more similar to blue than it is to pink. It is a similarity between the relationships among their respective members that make two property families counterpart families. Two properties of different property families are counterparts in virtue of occupying the same positions in their respective quality families. And a property's position is determined by the similarities and differences it bears to all of the other members of the quality family.

For instance, just as red is more similar to orange than it is to green, red* is more similar to orange* than it is to green*. Red* resembles and differs from the other colors* in ways that are homomorphic to the ways that red resembles and differs from the other colors. In virtue of this, red* and red occupy the same place in their respective quality families.

The homomorphism view also explains the correlations between the spatial properties of distal stimuli and the mental spatial properties of sensations.[3]

---

[2] I take physical colors to be sets of reflectance properties. Two surfaces that appear in normal lighting conditions to be the same shade of red are the same color. But such surfaces can have very different physical makeups. These surfaces, called metamers, pose a problem for the view that colors are physical light-reflectance properties of surfaces (see C.L. Hardin, 1993). But the problem can be countered by taking colors to be sets of reflectance properties, all of which yield the same ratio of light wavelength.

[3] This paper concerns the issue of location. But the homomorphism view accounts for other spatial properties, such as shapes and sizes. Squares are more similar to trapezoids than they are to circles. Likewise a square* sensation is more similar to a trapezoidal* sensation than

Sensations have apparent locations that normally correspond to the locations of distal stimuli. A red* sensation at the center of the visual field is normally caused by a red stimulus at the center of the distal visual field. The homomorphism view posits two distinct properties: Center-of-the-visual-field* and being at the center of the visual field. Stimuli in the center of the distal visual field normally cause sensations at-the-center-of-the-visual-field (CVF*, hereafter). A red* sensation to-the-left-of* a green* sensation is normally caused by a red stimulus to the left of a green stimulus, both of which are located in the distal visual field.[4]

The sum total of location* properties of visual sensations at a given time constitute the mental visual field at that time. So the CVF* is that location* equidistant* from all opposing points on the boundary* of the visual field, where the boundary* is defined by the limits of locations*. For instance, the left* boundary is set by the sensation to which no other sensation is to-the-left* of it.

Locations* within the mental visual field correspond to locations of entities in the distal visual field in virtue of resembling and differing from other locations* in ways homomorphic to the ways locations in the distal visual field resemble and differ from one another.

Two stimuli can resemble each other more than either resembles a third with respect to location in a distal visual field. Two objects to my left are more similar to each other than either is to an object to my right, with respect to at least one dimension of location. Both have the property of being to the left of me, while the third has the property of being to the right of me. The left objects will be more similar with respect to location to a fourth object directly in front of me than they will be to the object on the right. This is because being to the left of me is more similar to being directly in front of me than it is to being to the right of me (with respect to the left/right axis of location properties).

And sensations can resemble and differ with respect to mental location. Take a red* sensation off-to-the-left*, a yellow* sensation in the CVF*, and a blue* sensation off-to-the-right*. The red* sensation resembles the yellow* sensation more than it resembles the blue* sensation, with respect to location*. This is because to-the-left* is more similar to CVF* than it is to to-the-right*.

The structures of the quality families of the distal visual field locations and of the location* properties are homomorphic to one another. So CVF* and being in the center of the distal visual field are counterpart properties in virtue of their occupying the same place in their respective quality families, as fixed by the ways they resemble and differ from other properties of those families.

The result of this view is an explanation of how we see objects as being located where they are. We have visual sensations with location* properties. These sensations are not really located in any two-dimensional overlay visual field. Rather, the sensations are located* in the mental visual field. An afterimage appears where it appears because it has a certain location*. It appears to occlude the photographer's face because its being in that location* means that no sensations of his face can have that location* at that time.

Location* properties help explain how it is that having a CVF* sensation enables us to locate a distal stimulus directly in front of us. CVF* sensations carry information to the effect that there is something directly in front of one's eyes in virtue of CVF* being the counterpart property of being in the center of the distal visual field. It is in virtue of this counterpart relation that having a CVF* sensation helps us locate an object in the center of the visual field, as opposed to one off to the left of the visual field. And it is important to note that the homorphism view explains the counterpart relation in terms of similarity matrices that are readily accessible to us in ordinary visual experience and introspection.

When we introspect our sensations we pick them out, not by their perceptual role, but by their sensory qualities—that is, by their *-properties. When I introspect my sensation of a red patch to the left of another red patch, I pick out two sensations in virtue of their different locations*. That is, I pick them out in virtue of the ways they resemble and differ from one another and other sensations.

3. Clark's Rejection of Mental Space

Austen Clark (1996, 2000) rejects the existence of mental sensory fields. He claims that we need not mention locations of sensations to explain spatial experience. And he offers his feature-placing theory to this end.

Feature-placing aims to explain spatial experience in terms of the spatial properties of distal stimuli, the spatial properties of sensory receptors, and neural activation patterns. The only space needed is physical space.

According to Clark, "Sensing proceeds by picking out place-times and characterizing qualities that appear at those place-times." (2000, p. 74) A sensation identifies a location and qualifies it as being a certain way.[5] It does this in virtue of two variables. The

---

it is to a circular* sensation. The structure of the shape quality family and that of the shape* quality family are homomorphic to one another. Square* occupies the same position in its quality family as square occupies in its quality family.

[4] Stimulus location properties are determined relative to a perceiver. Which stimulus is to the left of another depends on the location from which one sees them.

[5] Stimulus location properties are determined relative to a perceiver. Which stimulus is to the left of another depends on the location from which one sees them.

sensation characterizes the place-time as being some way in virtue of its sensory qualities. A sensation qualifies a place-time as being red in virtue of the sensation's being red*. Clark's theory is in keeping with the homomorphism view with respect to so called secondary qualities.

But which place-time is qualified as being red is determined, not by some location* property, but by the firing of what Clark calls a sensory name—a stand-in for the mechanisms of spatial discrimination. These mechanisms identify place-times by what Clark calls place-coding, which he describes with respect to somesthetic experience.

A group of sensory receptors on the surface of the skin fire when stimulated, sending a neural impulse to the somatosensory cortex, where a certain neural activation pattern occurs. That neural activation pattern is the neural correlate of some bodily sensation—e.g., that of an itch (pp. 169-170). Where it is that the physical itch is felt to be depends on which groups of receptors fire (p. 173). These receptor groups are picked out by n-tuples of coordinates corresponding to the different dimensions in which the receptor groups vary in location. Similarly, the qualities of sensations can be coordinatized according to the dimensions in which those sensations vary (p. 176).

Accordingly, in a visual experience a surface is represented as being red in virtue of receptor groups on the retina firing in a certain way, leading to an activation in the visual cortex corresponding to red*. The red surface is represented as being off to the left in virtue of receptor groups on the left side of the retina firing. The red* state realized in the cortex is indexed to a particular place-coding n-tuple picking out that receptor group.

Having a sensation of red in the left of the visual field is a function of which receptor groups fire, and how they fire. The difference between a sensation of red in the left of the visual field and a sensation of red in the center of the visual field is just a difference in which retinal receptor groups fire. Cases of sensations without distal causes, and anomalies like afterimages are just misfirings of receptor pools. Feature-placing appears to have solved the problem of the spatial character of sensations without reference to mental space.

4. Why We Need Mental Space and Why Clark Does Too

But feature-placing cannot account for differences in the apparent locations of sensory qualities for two reasons. First, we cannot pick out our sensations without reference to some kind of spatial properties of those states. And second, we cannot identify the neurophysiological processes responsible for such variations without first individuating sensations by their mental spatial properties. Clark avoids positing mental space by resorting to spatial properties of stimuli and neurophysiological mechanisms. But his theory can only get off the ground if it accepts some sort of locations for sensations.

A visual sensation of a red patch to the left of a green patch is different from an experience of a red patch above a green patch. All of us who have visual experiences know this. We are conscious of these states as differing in some locational way. Without properties of sensations corresponding to spatial properties, we cannot discriminate between these two states. They would both be states of just a green patch and a red patch. So, unless the color* patches are located in physical space[6], and thus have spatial properties, they must either (a) be located in some sensory field, or (b) not exist. Clark must reject the existence of sensory states, or accept the existence of sensory fields and the sensory locations that comprise them.

But Clark does not reject sensations. He claims that sensations have sensory qualities such as red*, itchy* and high-C*. These are properties sensations must have if they fill the perceptual roles they fill. And they are properties in virtue of which we become conscious of those sensations when introspecting them. When I introspect the sensation I have when looking at a red firetruck, I am conscious of it as a red* sensation.

But I am conscious of such sensations as varying in apparent location as well. And two sensations can only help in discriminating between two differently located, but otherwise identical objects if those sensations differ in ways relating to the differences in object locations. The best explanation is that sensations have mental locations.

Further, we need to pick out our sensory experiences by their mental properties in order to identify their neurophysiological correlates. In order to identify the neurophysiological processes responsible for the appearance of a red patch in the center of my visual field, we need to pick out the appearance of a red patch in the center of my visual field. We do not do this by identifying properties of sensory receptors, nor of neural activation patterns. We pick out the appearance by its properties. Since such appearances can differ with respect to apparent location, there must be mental location properties.

Clark claims that variation in the spatial character of appearances is explained in terms of the spatial locations of distal stimuli and sensory receptors. Clark thus appeals to the properties of neurophysiological processes to explain spatial variations in experience.

But we determine that firings of receptor groups are responsible for certain locational features of sensory experience by discovering that those firings occur when

[6] Frank Jackson (1977) has argued that colors are mental entities that exist in physical space. An important distinction between his account of colors and Clark's account is that Jackson takes colors to be sense-data, whereas Clark takes mental colors to be properties of mental states.

and only when subjects have sensory experiences with those locational features. To determine which receptor groups are firing we monitor neurophysiological activity in the subjects. To determine the kind of sensory experience the subject is having we monitor the subject's overt and verbal behavior.[7] If the subject reports having a sensation of red in the center of the visual field, we infer that he has a sensation of red in the center of the visual field. These inferences rely on the presumed ability of the subject to pick out sensory experiences by their mental properties—in these cases, in part by their mental spatial properties. So Clark's explanation of the spatial character of experience relies on the existence of spatial qualities of sensations by which we pick out these sensations.

But the homomorphism avoids these problems. We pick out our sensations by their sensory qualities. These qualities include mental locations. And since we can pick out our sensations by these location* properties, we can come to identify the neurophysiological correlates of these states. Further, the homomorphism view explains the spatial character of visual experience without the implausible claim that such character depends on actual spatial locations of sensations and sensory qualities.

References

Boghossian, Paul & Velleman, David (1989) Colour as a Secondary Quality. In A. Byrne & D. Hilbert (Eds.) *Readings on Color, volume 1*. MIT.

Clark, Austen (1993) *Sensory Qualities*. Oxford University Press.

Clark, Austen (1996) Three Varieties of Visual Field. *Philosophical Psychology*, 9, no. 4, 477-495.

Clark, Austen (2000) *A Theory of Sentience*. Oxford University Press.

Dretske, Fred (1995) *Naturalizing the Mind.* MIT Press.

Hardin, C.L. (1993) *Color for Philosophers: Unweaving the Rainbow, expanded edition.* Hackett.

Harman, Gilbert (1990) The Intrinsic Quality of Experience. *Philosophical Perspectives*, 4, 31-52.

Harman, Gilbert (1996) Explaining Objective Color in Terms of Subjective Reactions. In A. Byrne & D. Hilbert (Eds.) *Readings on Color, volume 1*. MIT.

Jackson, Frank (1977) *Perception.* Cambridge.

Lewis, David (1972) Psychophysical and Theoretical Identifications. *Australasian Journal of Philosophy*, L, 3 (December), 249-258.

Peacocke, Christopher (1992) *A Study of Concepts*. MIT.

Rosenthal, David M. (1998) Sensory Quality and the Relocation Story. *Philosophical Topics*, 26, 1 and 2 (Spring and Fall), 321-350.

Rosenthal, David M. (1999) The Colors and Shapes of Visual Experience. In *Consciousness and Intentionality: Models and Modalities of Attribution*, Denis Fisette (Ed.). Dordrecht: Kluwer Academic Publishers.

Rosenthal, David M. (2000) Color, Mental Location, and the Visual Field. *Consciousness and Cognition* 9.

Sellars, Wilfrid (1956) Empiricism and the Philosophy of Mind. In W. Sellars, *Science, Perception and Reality.* Ridgeview, 1963.

Sellars, Wilfrid (1959) Being and Being Known. In W. Sellars, *Science, Perception and Reality.* Ridgeview, 1963.

Sellars, Wilfrid (1960) Phenomenalism. In W. Sellars, *Science, Perception and Reality.* Ridgeview, 1963.

Sellars, Wilfrid (1967) Sensibility and Understanding. In W. Sellars, *Science and Metaphysics*. Ridgeview, 1992.

Shoemaker, Sydney (1975) Functionalism and Qualia. *Philosophical Studies* XXVII, 5 (May), 292-315.

Tye, Michael (1995) *Ten Problems of Consciousness*. MIT Press.

Tye, Michael (1996) Perceptual Experience is a Many-Layered Thing. In E. Villanueva (Ed.) *Philosophical Issues*, 7, 1996, 117-126.

[7] I am sympathetic to David Lewis's (1972) account of psychophysical identifications. Lewis claims that identifying the neural correlates of mental states requires an understanding of the platitudes about mental states. This seems the right way to go about identifying the neural correlates of mental states inasmuch as mental states are the sensations, thoughts, desires, emotions, etc. that we report having. This should not preclude the importance of scientific psychology in making the identifications. Scientific psychology will have to draw distinctions that the folk do not always make themselves with respect to their mental states. These distinctions are drawn by running controlled experiments designed to isolate certain phenomena. But the psychologist can only design the experiments to draw these distinctions if he accepts that what he is trying to determine is somehow determined first by folk psychology.

HOW PRIMITIVE IS SELF-CONSCIOUSNESS?: AUTONOMOUS NONCONCEPTUAL CONTENT AND IMMUNITY TO ERROR THROUGH MISIDENTIFICATION

Roblin R. Meeks (rmeeks@gc.cuny.edu)
Department of Philosophy, The Graduate School and University Center,
The City University of New York, 365 Fifth Avenue
New York, NY 10016-4309 USA

Abstract

Traditionally, investigations into the nature of self-consciousness have focused on the peculiarities of the first-person pronoun. But can we extend the notion to non-language-using creatures as well, including pre-linguistic infants? José Luis Bermúdez has recently argued that creatures possessing no conceptual abilities whatsoever nevertheless possess states that can be considered primitive forms of self-consciousness. I discuss one such form Bermúdez gives—that of somatic proprioception—and show that it fails to satisfy the conditions he adopts for states funded by that type of perception to be representational as well as to be immune to error through misidentification. This conclusion forces a choice between abandoning either immunity to error through misidentification or a sharp conceptual/nonconceptual distinction with regard to representational states.

Introduction

Most traditional accounts of self-consciousness have focused exclusively on the peculiarities of the first-person pronoun. To be self-conscious from this perspective is to possess the ability to make judgments employing a first-person concept, judgments canonically expressed with 'I'. But do creatures lacking linguistic abilities thereby lack self-consciousness? After all, when hungry, even lobsters are self-possessed enough to avoid eating themselves. And what of pre-linguistic infants? If they are eventually to come to entertain thoughts involving a first-person concept, how does self-consciousness for them arise out of their wordless beginnings?

Venturing away from such traditional accounts requires that we should be clear concerning what we mean when we speak of a creature as self-conscious. In general, to be self-conscious, a creature must possess states with first-person content. We need to restrict our search further, however, for first-person content comes in (at least) two flavors. Consider the following examples:

(1) I am the winner of the New York Lottery.

(2) RM is the winner of the New York Lottery.

Intuitively it seems that (2) does not entail (1), for I can rationally believe that (2) is true while denying the truth of (1)—I could lack a further belief that I am identical with RM. In (1), I am thinking of myself nonaccidentally, perfectly aware to whom I am ascribing the property of lottery-winner, even if I have misread the numbers on my ticket and am actually no wealthier than before. In contrast, (2) leaves open the possibility that I am thinking of myself only accidentally, ascribing a property to someone unbeknownst to me who in fact turns out to be myself. Naturally, for me the above cases will further differ radically in the amount of joy expressed at their tokening. But the crucial distinction between the two illustrates the cardinal feature of self-consciousness: For a creature to be self-conscious it must be capable of possessing states that, like (1), have nonaccidental first-person content.

Can creatures lacking any conceptual resources whatsoever possess states that capture the distinction between (1) and (2), or at least approximate the nonaccidental nature of (1)? José Luis Bermúdez has offered an affirmative answer to this question, arguing at length in The Paradox of Self-Consciousness that certain forms of autonomous nonconceptual content—states with which a creature represents the world as being such-and-such a way despite possessing no conceptual resources whatsoever—can be considered forms of genuine self-consciousness.[1] We have good initial reason to agree with Bermúdez: Extending the range of types or forms of content that can correctly be characterized as genuinely first-personal gives us a hope of dispelling the mystery of how the richer,

[1] Bermúdez is motivated to look for nonconceptual content that is genuinely first-personal to escape what he calls the paradox of self-consciousness. This paradox is roughly that analyzing self-conscious thought solely in terms of a subject's mastering the first-person pronoun will rely upon the notion of him thinking of he himself as the author of the thought. Spelling out the "he himself" condition requires reference to the first-person pronoun, and we thus fall prey to circularity. Whether one finds Bermúdez's paradox compelling, it is an interesting question in its own right as to whether creatures lacking conceptual resources should be thought of as self-conscious and if so on what grounds.

conceptual forms of self-consciousness actually arise in the normal course of human psychological development.

In what follows we will consider one source of perceptual contents—namely somatic proprioception—that Bermúdez believes gives rise to genuine, albeit primitive, forms of self-consciousness. We will find, however, that a widely accepted condition that must be met for a state to be considered nonaccidentally first-personal stands at odds with certain nonconceptual states' being representational. In light of the incongruity, we face a choice between rejecting that condition, that nonaccidental first-person states be immune to error through misidentification, or accepting that a clear distinction between conceptual and nonconceptual states cannot be maintained.

Autonomous Nonconceptual Content

Elucidating exactly what nonconceptual content in general amounts to is a difficult task. Bermúdez himself is interested in establishing the existence of states with autonomous nonconceptual content to fend off circularity in a certain explanation of nonaccidental first-person thought. Though one can dispute his charge of circularity, his overall approach to primitive self-consciousness is instructive. He motivates the theoretical necessity of nonconceptual representational states via inference to the best explanation. Arguing on a broadly functionalist line, Bermúdez contends that no account of the behavior of an intentional system can be given without reference to representational states. However, certain intentional systems—including non-linguistic animals and pre-linguistic infants—lack concepts, yet still succeed, for example, in navigating their environment. We know that such creatures are representing their surroundings (and the states of their bodies) because no law-like relation holds between sensory input and behavioral output. Differences in behavior when faced with the same sensory input indicate that a creature is possibly misrepresenting a current state of the world or perhaps that its behavior is a function of a complex group of states, some of which differ from a previous occasion (a past predator can become prey, e.g.). Once general room has been made for states with autonomous nonconceptual content, Bermúdez goes to great lengths to provide specific examples of nonconceptual contents that qualify as primitive forms of self-consciousness.

One such example Bermúdez gives is that of somatic proprioception.[2] One's proprioceptive system provides a stream of information regarding the state of one's body, the position of limbs, skin and joint tension, bodily feedback during motion, etc. These states are representational states because they, like any other representational state, "serve as intermediaries between sensory input and behavioral output" (Bermúdez, 1998). Granting for the moment that such states are both representational and autonomously nonconceptual, how are we to determine if they qualify as forms of primitive self-consciousness? Bermúdez offers that such states must meet the two core requirements for any self-conscious thought: They must have immediate implications for action,[3] and they must be nonaccidentally about oneself. Skipping the former for the moment, thoughts are nonaccidentally about oneself, Bermúdez and many others argue, because they are immune to error through misidentification relative to the first-person pronoun. To assess the claim such states have to self-consciousness with any accuracy, we must briefly review what this condition amounts to more generally.

Immunity to Error Through Misidentification

In The Blue Book Wittgenstein (1958) distinguishes between what he calls 'I' used as subject and 'I' used as object. The latter, he claims, permits the possibility of misidentifying the referent of the first-person pronoun, whereas the former does not.[4] When uttering 'I am in pain'—the canonical instance of 'I' used as subject—Wittgenstein offers that the identification of the speaker is not in question: I cannot ascribe a felt pain to someone who, unbeknownst to me, is actually myself. In a genuinely self-conscious ascription of a property, it is no accident that I recognize that I am the subject of the ascription, for it could not be otherwise. In Wittgenstein's memorable phrase: "The man who cries out with pain, or says that he has pain, doesn't choose the mouth which says it" (Wittgenstein, 1958, emphasis his).

Sydney Shoemaker has done much work to elucidate and to extend this condition, labeling it with the now standard terminology "immunity to error through misidentification relative to the first-person pronoun" (Shoemaker, 1968).[5] For Shoemaker, roughly as for

[2] For a fairly extensive summary of the informational systems that constitute somatic proprioception, see the general introduction to Bermúdez, Marcel, & Eilan (1995).

[3] For a characterization of this requirement see Perry (1979).

[4] Indeed, Wittgenstein claims that 'I' in cases of its use as subject is not a referring expression at all. This position is endorsed and quite forcefully defended by Anscombe (1975).

[5] Shoemaker (1968) basically accepts Wittgenstein's distinction tout court, though he does hold that instances of 'I' in judgments immune to error through misidentification do genuinely refer. In recent work, Shoemaker (1994) has adopted Gareth Evans's (1982) coinage for this immunity, calling such judgments "identification free". The argument that follows does not depend on favoring a particular terminology, and therefore I will use the original phrase to avoid possible confusion. For a recent exploration of the kinds

Wittgenstein, a certain class of judgments permit error in the predicate position but do not leave the identity of the subject of the predication in question, for knowing in a particular way that a property is instantiated simply obviates the need for identifying its source. Bermúdez rightly points out, as Gareth Evans did before him, that these contents are immune to error through misidentification in virtue of the "evidence base from which they are derived, or the information on which they are based" (Bermúdez, 1998), not in virtue of any particular predicate or predicates. Ascriptions of pain to myself as well as to others employ the same predicate; the claim is that immunity issues from the way in which I know a pain to be present.[6] Fundamentally, Bermúdez– like nearly all other participants in this dialectic— accepts that contents cannot be considered genuinely self-conscious unless they possess this type of immunity.[7]

Somatic proprioception provides just such an evidence base, argues Bermúdez, for "somatic proprioception cannot give rise to thoughts that are accidentally about oneself" (Bermúdez, 1998). He writes:

> One of the distinctive features of somatic proprioception is that it is subserved by information channels that do not yield information about anybody's bodily properties except my own (just as introspection does not yield information about anybody's psychological properties except my own). It follows from the simple fact that I somatically proprioceive particular bodily properties and introspect particular psychological properties that those bodily and psychological properties are my own. (Bermúdez, 1998)

Focusing just on the particular bodily properties reported on by proprioception, how are we to assess the claim that I cannot be mistaken about within whose body those properties are instantiated when perceived in that way? For somatic proprioception to be a source of genuine self-consciousness, it must serve as an evidence base for contents where the subject cannot be in doubt, even for creatures lacking any conceptual resources whatsoever. Yet to qualify as representational– that is, to be considered contentful at all– thoughts funded by proprioception must allow for the possibility of misrepresentation. Misidentification is but a special case of misrepresentation, and hence endorsing immunity to error through misidentification at this primitive level precludes misrepresentation, which apparently serves to disqualify proprioceptive states from being representational.

To put the point another way, how can states funded by proprioception misrepresent? States in general can only "who" or "what" misrepresent– viz., they can misrepresent the subject of the state ("who") or the presence of a property ("what"), or presumably both. Misrepresentation of the "who" variety amounts to misidentification. To have "what" without "who" misrepresentation requires some representation of the subject with which a mistaken ascription can be made. Since nonconceptual states lack subject-predicate structure, no such representation of the subject is available in that case. Hence, to "what" misrepresent is to misidentify.

Unlike those who discuss immunity to misidentification as it relates to judgments, it is not at all clear that proponents of nonaccidental nonconceptual content have the philosophical machinery to relieve this tension. Evans, for example, does not fall into a similar predicament, for his 'I'-thoughts possess a conceptual structure that localizes— as Shoemaker's condition in its long form indicates— the immunity to error through misidentification relative to the first person pronoun. Misrepresentation can still occur with regard to the predicate position and the ascription of bodily properties, and hence immunity to misidentification and misrepresentation can co-exist in the same thought or judgment. Non-language-using creatures, of course, do not have the first-person pronoun at their disposal. Without conceptually structured thoughts, it seems that these types of subjects cannot possess contents that are both representational and immune to error through misidentification, for they have nothing that that immunity could be relative to.

Or do they? Bermúdez argues that inference to the best explanation warrants ascribing "protobeliefs", or nonconceptual belief analogs, to non-language-using creatures requiring intentional explanations to account for their behavior. As he presents them, perceptual protobeliefs[8] are nearly as rich as their conceptual correlates: they can embody "nonextensional modes of presentation" in terms of Gibsonian affordances, and they are somewhat compositional, though they do not allow for "global recombinability", failing to meet

---

of immunity, including fundamental ways in which Evans and Shoemaker disagree, see Pryor (1999).

[6] Cf. Evans (1982). Bermúdez also argues, persuasively I think, that Shoemaker's elucidation of immunity to error through misidentification should be stated in terms of justification as opposed to knowledge. For if one can still be mistaken about the instantiation of a predicate— even if one cannot be mistaken about the first-person identification in that case— that belief cannot be considered knowledge. It remains a question whether for Shoemaker this is possible.

[7] John Campbell (1999), for example, has recently remarked that "immunity to error through misidentification is a datum" that can be used to test the viability of various theoretical approaches to the first person.

[8] Bermúdez (1998) also briefly discusses instrumental protobeliefs, but our discussion can safely ignore them. Bermúdez draws this bit of his theoretical apparatus from Peacocke (1992).

Evans's Generality Constraint (Bermúdez, 1998; Evans, 1982). So structured, perceptual protobeliefs support primitive inference and the limited generation of further new nonconceptual contents from a set of others. Accordingly, perceptual protobeliefs so construed—including contents based on somatic proprioception—seem capable of supporting something like a discrete subject component, analogous to an 'I'-idea, that could serve as the locus of immunity to error through misidentification, as well as a predicative component that could misrepresent a property of the world or body.

One certainly becomes puzzled at this point, however. If nonconceptual contents based upon somatic proprioception can support both a component immune to misidentification and a component preserving the possibility of misrepresentation, then what are we to make of the original motivation for maintaining a clear conceptual/nonconceptual distinction with regard to contents? Indeed, it seems that inference to the best explanation warrants thinking of the constituents of protobeliefs as "protoconcepts". Much like concepts, protoconcepts could be defined in terms of their inferential role, where a protoconcept's inferential role can be cashed out in terms of the protopropositions or protobeliefs in which it features. As the analogy deepens between concepts and protoconcepts, we seem to have less reason to conclude that creatures lacking language likewise lack conceptual abilities of any sort, however limited or nascent. After all, the set of protopropositions may be quite limited for non-language using creatures, but they nevertheless succeed in satisfying two subtle and sophisticated philosophical criteria. Perhaps that success itself provides compelling evidence of some degree of concept possession.

Bermúdez himself would no doubt resist this approach since it seems to run afoul of what he calls the Priority Principle:

> The Priority Principle: Conceptual abilities are constitutively linked with linguistic abilities in such a way that conceptual abilities cannot be possessed by nonlinguistic creatures. (Bermúdez, 1998)

Priority was initially important because it "allows us to make a very clear distinction between conceptual and nonconceptual modes of content-bearing representation" (Bermúdez, 1998), and hence provides us with a means of explaining, for example, how conceptual forms of self-consciousness can arise over the course of normal human psychological development. Yet, given that protobeliefs are in some measure compositional and fund limited inference—indeed are constituted by protoconcepts—it is no longer clear how we can maintain a very clear distinction between conceptual and nonconceptual contents.

Still, perhaps the protoconcept/concept analogy runs fairly shallow, for even if non-language-using creatures possessed a range of protoconcepts defined in terms of protoconceptual roles, they do not have an explicit grasp of these roles. Such creatures are merely sensitive to the truth of inferential transitions. Bermúdez (1998) writes:

> Certainly, it is possible to be justified (or warranted) in making a certain inferential transition without being able to provide a justification (or warrant) for that inferential transition. It is a familiar epistemological point, after all, that there is a difference between being justified in holding a belief and justifying that belief. What does not seem to be true is that one can be justified in making an inferential transition even if one is not capable of providing any justifications at all for any inferential transitions. But providing justifications is a paradigmatically linguistic activity. Providing justifications is a matter of identifying and articulating the reasons for a given classification, inference, or judgment. It is because prelinguistic creatures are in principle incapable of providing such justifications that the priority thesis is true. Mere sensitivity to the truth of inferential transitions involving a given concept is not enough for possession of that concept. Rational sensitivity is required, and rational sensitivity comes only with language mastery.

For Bermúdez, then, possessing and deploying concepts demands a fairly advanced capacity to identify and to provide reasons for beliefs, and limited inferential ability—even an ability to make inferences that one is justified in making—does not indicate concept possession.

This seems a bit too stringent, however. Being able to give reasons as reasons is a function of possessing the concepts of justification, belief, and reason, among others. Imposing the further requirement on inferential ability that one recognize that one is in fact giving reasons may disqualify attributing conceptual abilities where we normally would be comfortable doing so. To take an example Bermúdez himself gives, the children in Susan Carey's experiments who concluded that a worm was more likely to have a spleen than a toy mechanical monkey are probably not in position to identify their reasons for this conclusion as reasons and to answer a call to justify their inferences. Still, he wants to credit these four-year olds with possessing the concepts HUMAN BEING, LIVING ANIMAL, INTERNAL ORGANS, and the inferential relations between them.

Conclusion

It seems that maintaining that nonconceptual contents be immune to error through misidentification entails that a sharp distinction between conceptual and nonconceptual contents must be abandoned. Perhaps we can spare a fairly strong distinction by instead abandoning the requirement that these contents be immune to error through misidentification. That is, we accept that protobeliefs are only minimally structured,

ultimately lacking the propositional precision required to support the weight of an immunity claim. It's not clear to me that we sacrifice much explanatory power in making this move, since we can still hold firmly to the second core condition for genuine self-conscious thought– namely, that nonconceptual proprioceptive contents must have immediate implications for action, which in fact they do (Bermúdez, 1998). Moreover, in preserving this second condition we still have a means of determining the class of nonconceptual contents that qualify as a form of genuine primitive self-consciousness. Alternatively, we can retain immunity to error as a necessary condition of self-consciousness, relinquishing instead the Priority Principle and the sharp conceptual/nonconceptual division that it was intended to capture. Choosing this route has interesting implications, for in doing so we greatly expand the range of creatures that can be said to possess conceptual capacities of one sort or another– including, evidently, those possessing some form of self concept.

Whatever route we choose, something, it seems, must be surrendered. For despite what doubts we might harbor concerning the lowly lobster, higher animals and our own infants should give us pause. Self-consciousness is certainly not ours alone; we just have yet to understand it in its more primitive forms.

Acknowledgements

I would like to thank Yen Chu, Bernard Kobes, Douglas Meehan, Erik Myin, Andrés Páez, David Rosenthal, Elizabeth Vlahos, and Joshua Weisberg for useful discussions and comments on an earlier version of this paper.

References

Anscombe, G. E. M. (1975). The first person. In S. Guttenplan (Ed.), Mind and language. Oxford: Clarendon Press.

Bermúdez, J. L. (1998). The paradox of self-consciousness. Cambridge, MA: MIT Press.

Bermúdez, J. L., Marcel A., & Eilan N. (Eds.) (1995). The body and the self. Cambridge, MA: MIT Press.

Campbell, J. (1999). Immunity to error through misidentification and the meaning of a referring term. Philosophical Topics, 26, 1999.

Cassam, Q. (1997). Self and world. Oxford: Clarendon Press.

Castañeda, H.-N. (1966). 'He': A study in the logic of self-consciousness. Ratio, 8, 130-157.

Evans, G. (1982). The varieties of reference. Oxford: Clarendon Press.

Peacocke, C. (1992). A study of concepts. Cambridge, MA: MIT Press.

Perry, J. (1979). The problem of the essential indexical. Noûs, 13, 3-21.

Pryor, J. (1999). Immunity to error through misidentification. Philosophical Topics, 26, 1999.

Shoemaker, S. (1968). Self-reference and self-awareness. The Journal of Philosophy, 65, 555-567.

Shoemaker, S. (1986). Introspection and the self. Reprinted in Shoemaker, 1996, 3-24.

Shoemaker, S. (1994). Self-knowledge and 'inner sense'. Philosophy and Phenomenological Research, 54, 249-314.

Shoemaker, S. (1996). The first-person perspective and other essays. Cambridge: Cambridge University Press.

Wittgenstein, L. (1958). The blue and brown books. Oxford: Basil Blackwell.

Automated Proof Planning for Instructional Design

Erica Melis (melis@dfki.de)
Christoph Glasmacher (christoph.glasmacher@cops.uni-sb.de)
Carsten Ullrich (cullrich@ags.uni-sb.de)
Peter Gerjets (pgerjet@gwdg.de)
DFKI Saarbrücken/Department of Psychology; Saarland University
D - 66041 Saarbrücken, Germany

Abstract

Automated theorem proving based on proof planning is a new and promising paradigm in the field of automated deduction. The idea is to use methods and heuristics as they are used by human mathematicians and encode this knowledge into so-called `methods`. Naturally, the question arises whether these `methods` can be beneficially used in *learning* mathematics too. This paper investigates and compares the effect of different instruction materials (textbook-based, example-based, and method-based) on problem solving performance. The results indicate that the performance for the method-based instruction derived from automated proof planning in the ΩMEGA system is superior to that of the other instructions that were derived from a textbook and an example-based classroom lesson. These results provide a first support for introducing proof planning based on methodological knowledge into the school curriculum for mathematics.

Introduction

Recent developments in automated deduction, one of the areas of Artificial Intelligence (AI), have shown the advantage of employing methods and heuristics used by human mathematicians. Naturally, the question is whether they can be beneficially used in teaching mathematics, for instance in interactive e-courses such as ACTIVEMATH (Melis et al., 2001).

The goal of the research reported in this paper has been to gather empirical evidence for the hypothesis that the knowledge we made explicit in proof planning `methods` for a restricted area of mathematics, namely limit problems, is indeed useful for learning to prove theorems in this area. A positive answer in this and other areas of mathematics can serve as a basis for the long-term goal to acquire `methods` to solve mathematical problems and then to use them to gradually change the teaching of mathematics.

To understand the interdisciplinary context, we will have a quick look at automated theorem proving.

Automated and Human Theorem Proving Traditional automated theorem proving systems such as OTTER have attained a remarkable strength in deductive search. They are, however, weak when it comes to nontrivial mathematical theorems where long range planning or other global search control is needed. Moreover, long proofs generated by these systems are almost incomprehensible. Therefore, techniques like proof planning that more closely follow the reasoning patterns observed in humans became more prominent.

The goal of automated proof planning (Bundy, 1988; Melis & Siekmann, 1999) is to identify and to employ human-like strategies and methods for theorem proving in order to avoid the almost exhaustive search in super-exponential search spaces that makes traditional automated theorem proving infeasible for most nontrivial mathematical conjectures. We investigated reports and mathematical textbooks (Melis, 1994) to make such strategies and methods explicit and then available for the ΩMEGA proof planner. Essentially, these methods are (generalized) macro-steps. This is in accordance with Koedinger and Anderson (1990) who investigated human theorem proving in geometry and found that humans employ macro-steps when proving theorems.

The identification and design of `methods` and control knowledge is very laborious as this kind of knowledge is not explicit in mathematical texts. However, some progress has now been made in the identification of mathematical methods and control knowledge (Melis, 1998). Based on these achievements we focus on questions such as

Is the knowledge that was made explicit for automated proof planning useful for supporting human learning of mathematical problem solving?

We are inclined to say *yes*. One reason is the explicit availability of this knowledge that can be used for proof presentation. An automated proof planner produces proof plans which in turn can be *presented in a more comprehensible way*. We investigated how proof presentation for teaching and learning can be generated from proof plans, see Melis and Leron (1999). Moreover, we investigated how such a presentation of proof plans can meet pedagogically and cognitively motivated requirements for presenting mathematical problem solutions and proofs, in particular the requirement for a hierarchically structured presentation originating from empirical results in Leron (1983) and Catrambone (1994).

A second reason is that this knowledge is needed for problem solving but not always present in textbooks (VanLehn, Jones, & Chi, 1992). Indeed, interviews with teachers of mathematics indicate a need for teaching methodological knowledge as captured in `methods`. Some even claim this is the essence of good teaching and

a source of improved learning and thus de-mystifying mathematics to some extent. As opposed to merely checking the correctness of single proof steps as in learning with traditional mathematical instruction, learning of *methods* should help in understanding the discovery of a proof. This leads to an improved performance based on understanding. The methodological knowledge includes the systematic construction of mathematical objects which is needed in many proofs.

The idea of making an expert's tacit problem solving knowledge explicit to learners is in accordance with some well known approaches in instructional psychology such as cognitive apprenticeship (Collins, Brown, & Newman, 1989) or the provision of instructional explanations (Chi, 1996).

Certainly, the success largely depends on the actual proof planning *methods* made explicit and encoded and therefore another direction of research, see Melis and Pollet (2000), aims at describing `methods` for interactive proof planning most appropriately. In addition to the evaluation of the concrete `methods` there is the more general question on whether the explicit teaching of relatively abstract methods helps in learning mathematics.

Although there are reasons to believe in instructional benefits, empirical evidence is required to substantiate the *yes*, and this is the focus of this report.

In this paper we present first empirical results. To begin with, proof planning is briefly reviewed, in particular proof planning of limit theorems which is the object of the described experiment.

Proof Planning Basics

Proof planning is based on classic AI-planning (Fikes & Nilsson, 1971) which reduces a goal to subgoals by introducing operators until all open subgoals match one of the initial state descriptions. When the sequence of operators is applied (in forward direction), the initial state is transformed into a state in which the goals hold. In proof planning, the goal is the theorem to be proved and the initial state consists of the proof assumptions.

For instance, for proving the theorem LIM+ which states that the limit of the sum of two real-valued functions f and g for a real number a is the sum of their limits L_1 and L_2, the conjecture to be proven is

$$\lim_{x \to a}(f(x) \quad g(x)) = L_1 \quad L_2$$

and the proof assumptions are

$$\lim_{x \to a} f(x) = L_1 \text{ and } \lim_{x \to a} g(x) = L_2.$$

A proof plan is a sequence of operators whose application realizes an inference from the proof assumptions to the theorem. In proof planning, the operators are called `methods`. They are frequently designed in a way corresponding to typical mathematical techniques such as `proof by induction`, `proof by refutation`, and `proof by diagonalization`, to quote some of the best-known `methods`. There are, however, less well-known `methods` which do not have a distinct name in mathematics. For instance, certain `estimation methods` for inequalities are typically not explicitly mentioned although they encode a frequently used trick. One of these `estimation methods` (`ComplexEstimate`) and another `method` (`TellCS`). These have been used in our experiments and are explained below.

Those non-name `methods` are often used only implicitly in course materials. This implicit treatment of proof `methods` is one reason why textbooks do not provide enough explanation of *how to find* a proof.

Proof Planning in the Limit Domain

In this section we describe a class of theorems, the way their proofs can be discovered mathematically, and the way proof planning in the ΩMEGA system implements this with `methods`.

The Theorems Limit theorems are taught at German high schools. Limit theorems claim something about the limit $\lim_{x \to a} f(x)$ for a function f or about the continuity of a function f.[1]

The definition of $\lim_{x \to a} f(x) = l$ describes formally that if x converges to a, then $f(x)$ converges to the limit l. The convergence $x \quad a$ means that the distance $x - a$ of x and a becomes arbitrary small. The definition of the limit describes that if $x - a$ becomes arbitrary small, then $f(x) - l$ becomes arbitrary small too. Put formally, for every arbitrary small real number ε exists a real number δ such that if $x - a < \delta$,[2] then $f(x) - l < \varepsilon$.[3]

Example Take the linear function $f(x) = 2 \quad x \quad 3$. When x converges to 0, then $f(x)$ converges to 3, i.e., for any arbitrary small ε, there is always a δ-environment $U_\delta(0)$ of $a = 0$ such that for any x in that environment $f(x)$ is in the ε-environment.

Counter Example Take as a counter example the function

$$f(x) = \begin{array}{rcl} 2 & : & x > 0 \\ -2 & : & x < 0 \end{array}$$

in Figure 1 which does not converge at point $x = 0$.

If ε is smaller than 2, there is always an x close to 0 for which $f(x)$ is not in the ε-environment of $l = \quad 2$ or of $l = -2$.

The Proofs The proofs of limit theorems have to suggest a δ, in relation to the given ε, such that the limit inequalities, e.g. $f(x) - l < \varepsilon$, hold. That is, a relation between ε and δ has to be determined such that for each x from the δ-environment of a the value $f(x)$ is in the ε-environment of l. Therefore, the standard proofs of these theorems are often called ε–δ–proofs.

Typically, textbooks postulate an appropriate relation between ε and δ out of the blue. Then they show

[1] or about the limit of a sequence which is a special case of a function.

[2] i.e., x is in the δ-environment $U_\delta(a)$ of a

[3] i.e., $f(x)$ is in the ε-environment $U_\varepsilon(l)$ of l

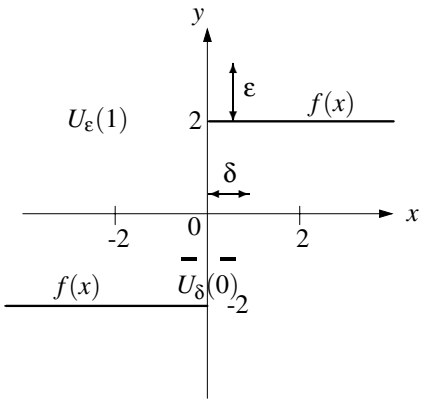


Figure 1: A function that does not converge at point $x = 0$

that the stipulated δ which is dependent on ε make the (inequality)-conjectures true. In contrast, proof *discovery* reveals the relation either by intuition or by systematically detecting conditions/constraints under which $f(x) - l$ becomes arbitrary small given that $x - a$ becomes arbitrary small. Those constraints result from analyzing the inequalities to be proven. This analysis often includes an abduction of new simpler inequalities/constraints sufficient to not invalidate the original ones.

These constraints may restrict the relation between ε and δ. For instance, if the constraints are $0 < \delta$ and $\delta < \varepsilon$, then $\delta = 2\ \varepsilon$ would be an invalid relation but $\delta = \frac{\varepsilon}{2}$ would be a valid one. When all possible constraints have been collected, then it is more transparent how to choose the relation between ε and δ. For instance, if the collected constraints are $0 < \delta$ and $\delta < \varepsilon$, then it is easy to see that the relation $\delta = \frac{\varepsilon}{2}$ satisfies the constraints. In particular, for complicated problems the systematicity is indispensable because ad hoc guesses and trial and error do not help much.

Proof Planning Proof planning for ε–δ–proofs (in a backward fashion) introduces a sequence of `methods` transforming $x - a\ < \delta$ to $f(x) - a\ < \varepsilon$:

$$\begin{array}{rrcl} & f(x) - l & < & \varepsilon \\ = & \ldots & < & \varepsilon \\ = & x - a & < & \delta. \end{array}$$

Each of the methods may yield restrictions on the relation of ε and δ. Therefore, proof planning systematically restricts the relation of ε and δ by uncovering constraints sufficient for making the inequalities true which are required in the theorem.

If a subgoal is a primitive inequality such as $0 < 1$ or $\delta < \varepsilon$, then `TellCS`[4] just collects it as a new constraint. If the constraints are not as immediate/primitive, then they can only be shown via a reduction to less complicated, primitive inequalities. For instance, to show $x^2 - a^2 < \varepsilon$ one might reduce the goal to the subgoals $x\ \ a < r$ and $x - a < \frac{\varepsilon}{r}$ for a number r to be determined and then conclude $x^2 - a^2 = (x\ \ a)\ \ (x - a) < r\ \ \frac{\varepsilon}{r} = \varepsilon$ and therefore $x^2 - a^2 < \varepsilon$. In proof planning such reductions are realized by estimation methods. One of those methods is `ComplexEstimate` whose simplified version is used in one of the instruction materials and described below.

Simplified ComplexEstimate The simplified `ComplexEstimate method` delivers the first reduction step in the following plan.

$$\begin{array}{rrcl} & f(x) - l & < & \varepsilon \\ = & k\ \ x - a & < & \varepsilon \\ = & x - a & < & \frac{\varepsilon}{k} \\ = & x - a & < & \delta \end{array}$$

It rewrites $f(x) - l$ to $k\ \ x - a$, determines the k which can be a number but also, in more complicated cases, a term like $x\ \ 1$ (see the Binomial computation above), and conjectures the subgoal that k has an upper bound (a real number r). The latter subgoal $k < r$ is a constraint and gives rise to establishing the relation $\delta = \frac{\varepsilon}{r}$ in order to guarantee $\delta < \frac{\varepsilon}{k}$ which implies the last proof step.

`ComplexEstimate`'s general procedure to determine k is polynomial division but manual computation may use simpler procedures in simpler cases, e.g. a Binomial formula.

This general `ComplexEstimate` (not used in the instruction materials) reduces an inequality goal to three subgoals (rather than two in the simplified version) by means of decomposing a term t into a linear combination $t = k\ \ a\ \ \ m$ for which an estimation of a is already known. It justifies the original goal by the three subgoals and the Triangle Inequality.[5] For difficult decompositions the `method` can call a polynomial division function without any problems.

The general `ComplexEstimate` as used in the automatic proof planner ΩMEGA covers the simpler cases for $k = 1$ and $m = 0$. Its generality allows for proving pretty complicated theorems that are beyond the range of our experiments. All test problems in the experiment require the special case $m = 0$ only. In the first, second, third, fourth, and fifth test problem, k is a real number, whereas in the sixth test problem k is the term $(x - 1)$.

Hypotheses

The overall goal of the study presented in this paper is an empirical validation of the assumption that the instructional presentation based on `methods` leads to an improved problem solving performance in mathematics. This differs from typical textbooks or classroom lessons where the methodological knowledge is currently not explicitly used.

The first hypothesis states that instructional material that includes information about proof-generation methods improves the overall problem solving performance.

[4] for "Tell the Constraint Solver".

[5] $A\ \ B < A\ \ \ B$

The second hypothesis postulates that the method-based instruction is especially helpful in solving far-transfer test problems that presuppose the generation of new solution paths.

To test the first hypothesis instructional material based on ΩMEGA's proof plan `methods` was designed. The method-based instructions were contrasted with conventional instruction materials: textbook-based instruction and example-based instruction.

To test the second hypothesis test problems of different transfer distance were used[6] .

Experiment

Method

Participants The subjects were 38 students of Saarland University, Germany who either participated for course credit or payment. Average age was 24.1 years.

Materials and procedure Each student was provided with the following material in a booklet: (1) An introduction that described the nature and purpose of limits. Additionally, the introduction presented a definition of the notion of an *environment* as a prerequisite for the formal definition of *limit*. (2) A formal definition of the notion *limit* together with an illustrating graph. (3) One worked-out example that illustrated how the limit $\lim_{x \to a} f(x)$ for a given function f and a given value a can be proven. Depending on the experimental conditions different solution approaches were selected in the worked-out examples.

Subjects were advised to study the instructional material carefully. After reading the booklet subjects had to solve six test problems that differed in their transfer distance with respect to the instructional example. The six text problems were of increasing difficulty and decreasing structural similarity to the example explained in the instruction.

Design and dependent measures Four different instructional materials were designed as independent variables: Textbook-based instruction, example-based instruction, and two types of method-based instruction (only differing in the sequence of the parts of instructional materials). The instructional conditions differed only with respect to the solution approach for the worked-out example and with respect to the sequence of the instructional materials.

In the textbook-based instruction the introductory page was immediately followed by a short formal definition of the notion *limit* and an illustrating graph. Subsequently, one example of a worked-out ε-δ-proof for a linear function ($f(x) = x \quad 2$, with x being undefined at $x = 1$) was presented. The example solution was taken from an university-level textbook. The textbook-based instruction merely postulated the pivotal relation between ε and δ without derivation from more general principles. The mere stipulation of pivotal assumptions is a frequent feature of example proofs in textbooks.

The example-based instruction differed from the textbook-instruction in that the example problem was presented immediately after the introductory page. To establish a general relation between ε and δ, suitable values for δ are introduced for several concrete ε values of decreasing size. This approach allowed for an inductive derivation of a general relation between these two parameters. Additionally, the example-based instruction differed from the textbook-instruction in the sequence of the instructional materials: The example proof was presented before the formal definition of the notion *limit* and the respective illustrating graph were introduced.

The method-based instruction took the `methods` simplified `ComplexEstimate` and `TellCS` from ΩMEGA's proof planner and described an example solution explicitly using `ComplexEstimate` and `TellCS` (the collection of constraints). It shows how `ComplexEstimate` reduces a complicated estimation to several simpler ones. As a general approach it also employs the collected constraints for defining a relation between ε and δ. The `methods` are applied to prove the example problem and an abstract description of the `method` is provided.

Two versions of this method-based instruction were designed that differ with respect to the sequence of instructional materials. In version A the definition of the notion *limit* was followed by an abstract description of `ComplexEstimate` and an illustrating example applying this `method`. In version B the worked-out example was presented before the notion *limit* was defined and the `ComplexEstimate` method was described in a more abstract way.

As dependent variables problem-solving time and problem solving performance for the six test problems were registered. The test problems differed in transfer distance. The first two test problems were isomorphic to the example used in the instructional material (proving a limit for a linear function of the form $f(x) = x \quad b$). The next three test problems were near-transfer problems (proving a limit for a linear function of the form $f(x) = ax \quad b$). Finally, a far-transfer test problem had to be solved (proving a limit for a quadratic function of the form $f(x) = ax^2 \quad bx \quad c$). After the experiment, data were collected by means of a questionnaire, in particular, the subjects' last maths grade in school, the subjects' interest in mathematics, sociodemographic data, and whether they were taught anything about limit theorems in (past) school lessons.

Results

The six test-problem solutions were scored as follows. For a totally correct answer a score of 1 for isomorphic problems, a score of 2 for near-transfer problems, and a score of 4 for far-transfer problems was assigned. Hence, the maximum total score is 12. 50% of the full score were assigned to a solution, if the answer was correct except for minor, nonconceptual mistakes (e.g. numerical calculation errors, mixing up δ and ε in the solution

[6] Transfer distance is a measure for structural similarity.

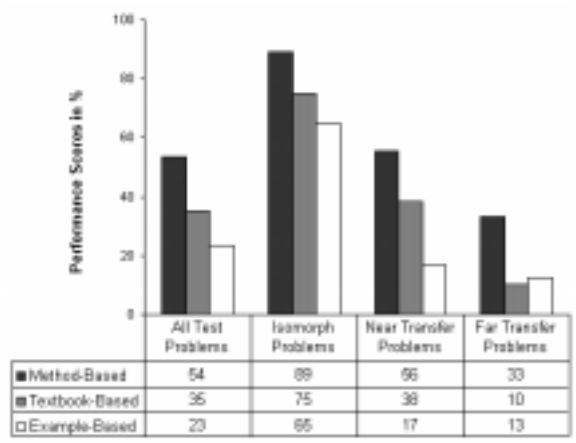


Figure 2: Mean performance scores (in percentage of possible maximum score) as a function of instructional condition and transfer distance between test problems and example problems

equation). 75% of the full score were assigned in case of incorrect solution of the polynomial in the last test problem.

Nonparametric tests were used in all performance analysis because of distorted distributions. In a first step, we compared the two method-based instructions with respect to performance differences. Mann-Whitney U-tests revealed that there were no differences in the total problem-solving score (U(9, 9) = 36; p(two-tailed) = .69) or in problem-solving time (U(9, 9) = 39; p(two-tailed) = .89). Thus, both method-based instructions were collapsed for further analysis. An overall comparison of the method-based, the example-based and the textbook-based instruction with Kruskal-Wallis H-test revealed that there were significant differences in the total problem-solving score ($\chi^2(2, N = 38) = 5.87$; $p = .05$) but not in problem-solving time ($\chi^2(2, N = 38) = 2.45$; $p = .29$). The instructional conditions did not differ with respect to the last math grade in school, domain-specific knowledge they were taught in school, interest in mathematics, sex, and age. Figure 2 provides the mean performance scores (in percentage of possible maximum score) for all three instructional conditions and all levels of transfer distance.

Paired one-tailed comparisons with Mann-Whitney U-tests (see Table 1) yielded that the method-based instruction outperformed the textbook-based instruction (marginally) as well as the example-based instruction with respect to the total problem-solving score. The textbook-based instruction and the example-based instruction did not differ in total problem-solving score.

A more detailed analysis revealed that the method-based instruction and the textbook-based instruction differed marginally with respect to isomorphic problems and to far-transfer problems but not with respect to near-transfer problems. The method-based instruction and the example-based instruction differed with respect to all performance measures, at least marginally. The textbook-based instruction and the example-based instruction did not differ with respect to isomorphic problems and far-transfer problems. However, there was a marginal significant difference with respect to near-transfer problems.

Table 1: Comparison between all instructional conditions with respect to all levels of transfer distance (one-tailed Mann-Whitney U-tests)

	All Test Problems	Isomorph Problems	Near Transfer Problems	Far Transfer Problems
Method (n_1=18) vs. Textbook (n_2=10)	**p = 0,08** U = 60,5	**p = 0,10** U = 69,5	p = 0,17 U = 71,5	**p = 0,07** U = 66
Method (n_1=18) vs. Example (n_3=10)	**p = 0,01** U = 43,5	**p = 0,07** U = 66,5	**p = 0,01** U = 48	**p = 0,10** U = 68
Textbook (n_2=10) vs. Example (n_3=10)	p = 0,15 U = 36,5	p = 0,37 U = 46	**p = 0,10** U = 35,5	p = 0,31 U = 46

Discussion

As postulated in our first hypothesis the method-based instructional material based on ΩMEGA's proof plan presentation has a significant beneficial effect on learners' subsequent problem-solving performance. Compared to more conventional instructional formats usually found in textbooks and highschool lessons the method-based instruction improves learners' problem-solving performance without requiring more time to be invested.

Contrary to the expectation expressed in our second hypothesis, the performance improvements due to the method-based instructional format are not larger for far-transfer test problems than for isomorphic and near-transfer test problems. An explanation for this unexpected result might be that the far-transfer test problem has been chosen as a too-far one that requires an additional computation (polynomial division) the subjects might have been not capable to carry out or did not even try.

To conclude, the results indicate that the method-based instruction that originated from proof planning *methods* implemented in ΩMEGA is superior to the two other instructions in terms of subsequent problem solving performance. These results provide first evidence that proof planning based on mathematical knowledge may also be used and introduced into highschool curricula for mathematics.

Conclusion

Is the methodological knowledge used in proof planning useful for human learning of maths problem solving?
The results of our experiments indicate that the method-based instruction that originated from automated proof planning is, indeed, superior to the two other instructions in terms of subsequent performance. These results provide first support for introducing proof planning based on

methodological knowledge into the highschool curricula for mathematics.

It is not necessary to restrict this methodological knowledge to `methods` which have been acquired for and used in automated proof planning. We can, however, re-use these results. Then the advantage is that those `methods` are formalized and implemented and, therefore, can be employed by a system supporting *interactive* problem solving.

The presented empirical results are limited, however, to only one area of highschool mathematics. Future work will try to provide similar evidence in other areas as well.

Interestingly, we met many committed mathematics teachers in Germany who have been engaged in activities targeting a similar idea without knowing, of course, about automated theorem proving and proof planning. Their concern is a reshaping of mathematics lessons that aims at learning problem solving methods, heuristics, and structuring problems and solutions rather than at memorizing facts and procedures.

Future Work

In the future we will replicate the experiment reported here with several augmentations. First, we will obtain think-aloud protocols to get more detailed insights into the learning and problem-solving processes elicited by different instructional materials. Second, we will try to shed more light on the results with respect to second hypothesis by adjusting the difficulty of the far transfer test problems. Third, we will additionally consider certain features of instructional situations like domain-specific prior knowledge or degree of time pressure that have been shown to influence the profitability of different instructional materials (Gerjets, Scheiter, & Tack, 2000).

Another line of research will pertain to the fact that the provision of profitable instructional materials does not ensure that learners indeed use these materials appropriately. This is especially true for computer-based learning environments that allow learners to control for many aspects of the learning process, e.g. the selection of instructional materials (Gerjets, Scheiter, & Tack, 2001). Therefore, we will examine whether learners select method-based instructional materials when they are allowed to choose between different types of information in electronic learning environments. Finally, we will design experiments investigating the influence of explicitely teaching control knowledge (i.e. knowledge on when to choose which method) in addition to the teaching of method knowldege.

References

Bundy, A. (1988). The use of explicit plans to guide inductive proofs. In E. Lusk & R. Overbeek (Eds.), *Proc. 9th international conference on automated deduction (cade-9)* (Vol. 310, p. 111-120). Argonne: Springer.

Catrambone, R. (1994). Improving examples to improve transfer to novel problems. *Memory & Cognition*, *22*(5), 606-615.

Chi, M. (1996). Constructing self-explanations and scaffolded explanations in tutoring. *Applied Cognitive Psychology*, *10*, 33-49.

Collins, A., Brown, J., & Newman, S. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of robert glaser* (p. 453-494). Hillsdale, NJ: Lawrence Erlbaum Associates.

Fikes, R., & Nilsson, N. (1971). STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, *2*, 189-208.

Gerjets, P., Scheiter, K., & Tack, W. (2000). Resource-adaptive selection of strategies in learning from worked-out examples. In L. Gleitman & A. Joshi (Eds.), *Proc. of the annual conference of the cognitive science society* (p. 166-171). Mahwah, NJ: Erlbaum.

Gerjets, P., Scheiter, K., & Tack, W. (2001). *Problems of example selection and example processing in hypertext-based learning environments* (Tech. Rep.). Saarbr"ucken, Germany: Saarland University.

Koedinger, K., & Anderson, J. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science*, *14*, 511-550.

Leron, U. (1983). Structuring mathematical proofs. *The American Mathematical Monthly*, *90*, 174-185.

Melis, E. (1994). How mathematicians prove theorems. In *Proc. of the annual conference of the cognitive science society* (p. 624-628). Atlanta, Georgia U.S.A.

Melis, E. (1998). The "limit" domain. In R. Simmons, M. Veloso, & S. Smith (Eds.), *Proc. of the fourth international conference on artificial intelligence in planning systems* (p. 199-206).

Melis, E., Andres, E., Goguadse, G., Libbrecht, P., Pollet, M., & Ullrich, C. (2001). Activemath: System description. In *Proc. of the international conference on artificial intelligence in education.*

Melis, E., & Leron, U. (1999). A proof presentation suitable for teaching proofs. In S. Lajoie & M. Vivet (Eds.), *international conference on artificial intelligence in education* (p. 483-490). Le Mans: IOS Press.

Melis, E., & Pollet, M. (2000). Domain knowledge for search heuristics in proof planning. In *Aips 2000 workshop: Analyzing and exploiting domain knowledge* (p. 12-15).

Melis, E., & Siekmann, J. (1999). Knowledge-based proof planning. *Artificial Intelligence*, *115*(1), 65-105.

VanLehn, K., Jones, R., & Chi, M. (1992). A model of self-explanation effect. *Journal of learning science*, *2*(1), 1-59.

Modeling an Opportunistic Strategy for Information Navigation

Craig S. Miller (cmiller@cs.depaul.edu)
School of Computer Science, DePaul University, 243 S. Wabash Ave.
Chicago, IL 60604 USA

Roger W. Remington (rremington@mail.arc.nasa.gov)
NASA Ames Research Center, MS 262-4
Moffet Field, CA 94036 USA

Abstract

A computational model of a user navigating Web pages was used to identify factors that affect Web site usability. The model approximates a typical user searching for specified target information in architectures of varying menu depth. Search strategies, link ambiguity, and memory capacity were varied and model predictions compared to human user data. A good fit to observed data was obtained for a model that assumed users 1) used little memory capacity; 2) selected a link whenever its perceived likelihood of success exceeded a threshold; and, 3) opportunistically searched below threshold links on selected pages prior to returning to the parent page.

Introduction

The World Wide Web continues to revolutionize how people obtain information, buy products, and conduct business transactions. Yet many companies and organizations struggle to design Web sites that customers can easily navigate to find products or information. The identification of factors that affect the usability of the World Wide Web has become increasingly important. While many of these factors concern the graphical layout of each page in a Web site, the way in which the pages link to each other, often called the site's information architecture, plays a decisive role in the site's usability, especially for sites allowing access to large databases (Rosenfeld & Morville, 1998) Our effort focuses on understanding how a site's information architecture impacts a user's ability to effectively find content in a linked information structure such as a Web site.

We develop our understanding through the construction and testing of a working computational model. The model simulates a user navigating through a site making choices about whether to select a given link or evaluate an alternate link on the same page. Constructing and testing a working model not only complements empirical studies, but also offers advantages over empirical usability testing. Empirical studies are generally too expensive and time consuming to address the wide range of content, configurations, and user strategies that characterize the Web. In contrast, an implemented model can run thousands of simulated sessions in minutes. Also, empirical studies do not inherently provide explanations for their results and thus make it more difficult to determine how a given result generalizes to other circumstances, whereas a cognitive model can describe the underlying processes that produce behavior. For example, computational models have been used to highlight patterns of interactions with a browser (Peck & John, 1992) and report on the accessibility of the site's content (Lynch, Palmiter & Tilt, 1999).

In this paper, we build upon methods that we presented in an earlier paper (Miller & Remington, 2000a). For the sake of presentation, we describe the methods and our model in its entirety. We introduce a new navigation strategy and show how the model's aggregate behavior tightly fits results from an empirical comparison of different site architectures (Larson & Czerwinksi, 1998). Finally, we experiment with the model's assumptions by exploring alternate designs and parameters in order to help identify critical elements in the model's design.

Modeling Information Navigation

We simulate common patterns of user interaction with a Web site with the goal of providing useful usability comparisons between different site architectures. A model that precisely replicates a user's navigation is not possible, nor is it necessary. Useful information can be obtained from a simple model that captures functionally significant properties of the user and site architecture. Here we show how a simple model can predict and explain benefits of one design over another, such as when it is advantageous to use a two-tiered site instead of a three-tiered site.

In constructing our model, we use the following principles:

- The model should only perform operations that are within the physical and cognitive limitations of a human user. In Web navigation, for example, limits on visual attention dictate that a user can only focus upon (and evaluate) one link at a time. Likewise, limits on short-term memory dictate

navigation strategies that minimize memory requirements, an assumption consistent with evidence that people often adopt memory minimization strategies (Ballard, Heyhoe, Pook, & Rao, 1997).

- The model should make simplifying assumptions whenever they are not likely to have much impact on aggregate behavior. For example the model takes a fixed amount of time to evaluate a link even though human users' times are certainly variable. Since the model simulates the average user, this simplification will provide a good fit given a reasonable estimate of fixed time from human performance data (Card, Moran & Newell, 1983).
- The model should employ the most effective strategy for a given environment unless compelling evidence from human usage suggests otherwise. Given the large set of navigation strategies that can operate within reasonable physical and cognitive limitations, we examine a strategy that is most effective within known cognitive constraints. This design constraint is the rationality principle (see Card, Moran & Newell, 1983), which assumes that human cognition is generally rational.

Representing a Web Site

Our model interacts with a simplified, abstract representation of a Web browser and a Web site. Each site has one root node (i.e. the top page) consisting of a list of labeled links, each leading to a separate child page. For a shallow, one-level site, child pages are terminal pages, one of which contains the target information that the user is seeking. For deeper, multi-level sites, a child page consists of a list of links, each leading to child pages at the next level. The bottom level of all our sites consists exclusively of terminal pages, one of which is the target page. Our examples are balanced trees (i.e. pages at the same level have the same number of links) since we generally compare our results to studies that use balanced tree structures (e.g. Miller, 1981 and Larson & Czerwinski, 1998). However, our representation does not prevent us from running simulations on unbalanced trees, or even on structures involving multiple links to the same page and links back to parent pages.

When navigating through a site, a user must perceive link labels and gauge their relevance to the targeted information. Rather than model the complex and interesting process of link evaluation, we fix a number for each link, which represents the user's immediately perceived likelihood that the target will be found by pursuing this link. This simplification allows us to easily investigate a range of numerical relationships between the link label and the target information.

In an ideal situation, after evaluating a link, the user would know with certainty whether to select and pursue that link. Figure 1 represents a site with such links. Each link (underlined number) on each Web page is understood without ambiguity. The user need only follow the links labeled with a 1.0 to find the targeted page with no backtracking. We describe the architecture of this site as having a two-tiered, 4x2 structure.

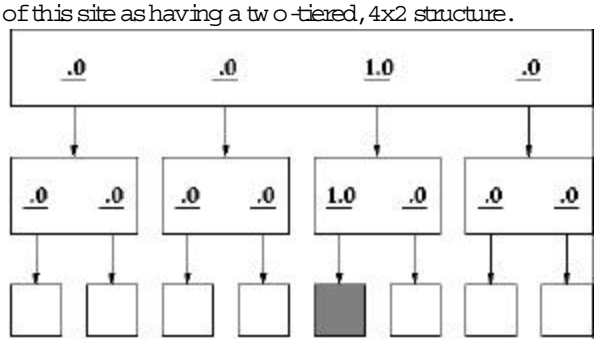


Figure 1 Site with sure path

More often, the user is less certain of which link to select. The links in the site shown in Figure 2 are more ambiguous. For the top page, the most likely link has a perceived likelihood of only .7, thus indicating that the user is less certain that this link will lead to the targeted item. In some cases, a user strategy that merely follows the most likely links would directly lead to the target. However, this figure shows a site where the user would find the target under what he or she perceives as a less plausible sequence of link selections (the target is under a likelihood value of 0.2 instead of the 0.5 value). In this way it is possible to represent sites that differ widely in strength of association between link label and target information.

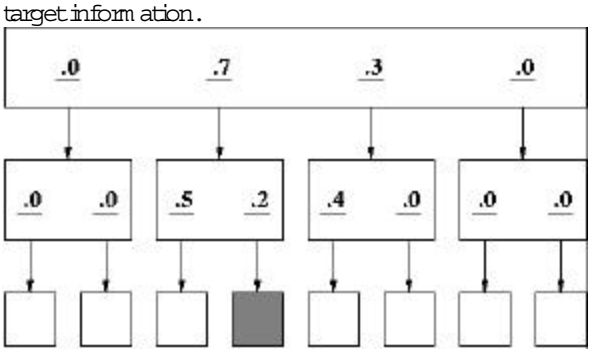


Figure 2 Site with ambiguous labels

Modeling the Browser and User Actions

Byrne et al. (1999) found that selecting a link and pressing the Back button accounted for over 80% of the actions used for going to a new page. Consequently, we focused our modeling on the component actions underlying these behaviors. These include:

- Selecting a link
- Pressing the Back Button

- Attending to and identifying a new page
- Checking a link and evaluating its likelihood

To further simplify our model, attending to and identifying a new page can be folded into the actions of Selecting a Link and Pressing the Back Button since this action only occurs when either of these actions occur. Our revised model has three primitive actions:

- Selecting a link (and attending to and identifying a new page)
- Pressing the Back Button (and attending to and identifying a new page)
- Checking a link and evaluating its likelihood

Because of physical and cognitive limitations, only one of these actions can be performed at any one time. Fixed times are assigned to each action to account for its duration during a simulation. The model also simulates changing the color of a link when it is selected so that the modeled user can "perceive" whether the page under this link was previously visited.

Modeling Navigation Strategies

The model navigates a Web site by serially executing these three primitive actions, meaning that links are sequentially evaluated. Serial evaluation is motivated by evidence that the human user has a single unique focus of attention that must be directed at the link for this decision.

A user could employ any of a large set of possible strategies for link selection that vary in sophistication. Two examples include:

- The threshold strategy: The user immediately selects and pursues any link whose probability of success exceeds threshold.
- The comparison strategy: The user first evaluates a set of links and then selects the most likely of the set.

The threshold strategy is most effective if the first likely link actually leads to the targeted object. The comparison strategy is more effective only if a likely link is followed by an even more likely link that actually leads to the targeted item. Both represent simple yet effective strategies. We chose to begin by examining the threshold strategy on the principle that it requires the fewest computational (cognitive) resources.

The model is neutral as to the actual order in which the links are evaluated. The design and layout of a page principally determine which links a user would evaluate first. Any understanding of how page layout and design affect the user's focus could eventually be incorporated into the model. For our purpose of investigating site structure, the model simply establishes a fixed order in which links are evaluated for each run. To avoid systematic order biases, our simulations randomly place the targeted item at a different terminal page for each run.

With the appearance of a new page, the model's threshold strategy first attends to the page, which, if it is a terminal page, includes checking if it contains the target information. If it does not, the model sequentially scans the links on a page selecting any link whose likelihood is equal to or above a fixed threshold (0.5 in the simulations reported below). When a page appears by selecting a link, the process of checking and scanning the page is repeated.

Once the model detects no unselected links above the threshold value, it returns to the parent page by pressing the Back button and continues scanning links on the parent page starting at the last selected link. It does not scan links it has already evaluated. Determining the last link selected places no demands on memory since the last selected link is easily detected by its color, and many browsers return the user to the location of the last selected link.

Selecting the most probable link will often lead to the targeted item. However, sometimes the targeted item lies behind ostensibly improbable links and, after some initial failures, human users must start selecting links even if the link labels indicate that they will probably not lead to the targeted item. An earlier version of our model (Miller & Remington, 2000a) started selecting improbable links only after completing a full traversal of the site. We will call this the traverse-first strategy. However, a more effective strategy may opportunistically select improbable links at a lower tier immediately after trying the more probable links and before returning to a higher tier in the site. We call this the opportunistic strategy.

Figure 2 illustrates how the opportunistic strategy may be more effective. The model scans across the top page and selects the second link (0.7). On the second level it selects the first link it encounters (0.5). After discovering that this is not the targeted item, it returns to the page on the second level. However, before returning to the top level, it temporarily reduces its threshold to 0.1, selects the second link (0.2) and finds the target on the new page. Had the targeted item been elsewhere in the site, the strategy backs up twice, returning to the top-level and resetting the threshold to the previous value (0.5).

The opportunistic strategy is more efficient than the traverse-first strategy. First, it explores less probable links when the cost of doing so is minimal, that is, when the less probable links are immediately available. Secondly, it implicitly takes into account the positive evaluation of the parent link, which had indicated a high likelihood that the targeted item is under a link on the current page.

The opportunistic strategy is not efficient if employed in cases where all the links on a page have very low likelihood values (defined as less than 0.1). In such cases our model assumes that the user has sufficient memory to know that rescanning the page would be futile, and returns to the parent page. This memory of knowing that the page has nothing worthwhile only lasts as long as the model remains on the current page. Thus, if the model leaves the page and then returns to this same page, the model must assume that the page may be worth rescanning and the opportunistic strategy is employed. This qualification is also consistent with our design principles in that it contributes to an effective strategy while minimizing memory resources.

While generally consistent with our design principle of preferring strategies that place minimal demands on memory, the opportunistic strategy does require state values to be held in memory. If opportunistic search fails to find the targeted item, the model must reset the link selection threshold to the previous value upon returning to the upper level. Resetting a value requires storing the old value before reducing it. Storing and recalling one or two values reasonably fall within the limits of human cognition, but storing and recalling an arbitrary number of values does not. For this reason, our model allows us to fix a limit on the number of previous threshold values it can recall. We initially set this number to one, but later in this paper we will explore the impact of being able to store and recall additional values.

Part of our reason for adopting the opportunistic strategy in place of the traverse-first strategy was our examination of usage logs for a site search task. We conducted a pilot study using a Web site whose structure mirrored a popular department store's organization. Preliminary results suggest that users frequently select ostensibly less probable links before backtracking to other possibilities (see Miller & Remington, 2000b, for more details and an example). We plan future studies that could further identify usages of this strategy.

Simulation Parameters

Our previous work established plausible time constants for link evaluation and link selection (Miller & Remington, 2000a). We compared the model and results from hierarchical menu selection studies and obtained good fits with link evaluation costs set to 250 ms and link selection costs set to 500 ms. The use of time constants is well established (e.g., Card, Moran, & Newell, 1983) and these values are consistent those previous estimates.

To assign likelihood factors to the links, the ideal link values (1, 0) are perturbed with noise according to the formula below:

$$g * n + v$$

where g is a number chosen randomly from a standard normal gaussian distribution (mean=0, stdev=1); n is the noise factor multiplier (equivalent to increasing the variance of the normal distribution); and v is the original likelihood value (0 or 1). Since this formula occasionally produces a number outside the range from zero to one, our algorithm may repeatedly invoke the formula for a link until it generates a number in this range. The noise factor n thus models the level of label ambiguity in the site. Higher levels of ambiguity lead to more frequent backtracking, which may be more prominent in Web search than menu search.

Simulations

To further evaluate the model's design decisions, we compare its performance to the Web navigation results of Larson and Czerwinski (1998). They studied users navigating two-tiered (16x32 and 32x16) and three-tiered (8x8x8) site architectures that were otherwise comparable. Participants took significantly longer to find items in the three-tiered site (58 seconds on average) than the two-tiered sites (36 seconds for the 16x32 site and 46 seconds for the 32x16 site).

Simulations of the Opportunistic Strategy

For our simulations using the opportunistic strategy, sites were constructed as described above, except that the noise was not applied to the bottom level, which leads to the terminal pages. This reflects the fact the participants in Larson & Czerwinski could clearly tell whether the link's label matched the text of the targeted item.

For each site architecture (8x8x8, 16x32, and 32x16) 10,000 simulations were run using the following time costs: 250ms for evaluating a link, 500ms for selecting a link, and 500ms for return to the previous page (pressing the back button). Following Larson and Czerwinski (1998), any run lasting more than 300 seconds was coded as lasting 300 seconds.

Figure 3 shows the calculated mean times of the simulation runs. Not surprisingly, the time needed to find a target increased with link ambiguity. What is more interesting is how link ambiguity interacts with site structure. The 8x8x8 architecture produced slightly faster times at low levels of noise but substantially slower times at noise levels above 0.2. At these higher noise levels the results are consistent with the human users. At noise levels of 0.4 and higher, simulated times were faster with the 16x32 architecture than the 32x16 architecture. This difference was also noted in the study with human users, albeit not reported as statistically significant.

At a noise level of 0.4, the simulation results closely match the human results in absolute terms: 62s

(compare to 58s for humans) for 8x8x8, 43s (compare to 46s) for 32x16, and 35s (compare to 36s). It appears that the 0.4 serves a good parameter estimate describing the amount of label ambiguity in the sites used by Larson and Czerwinski.

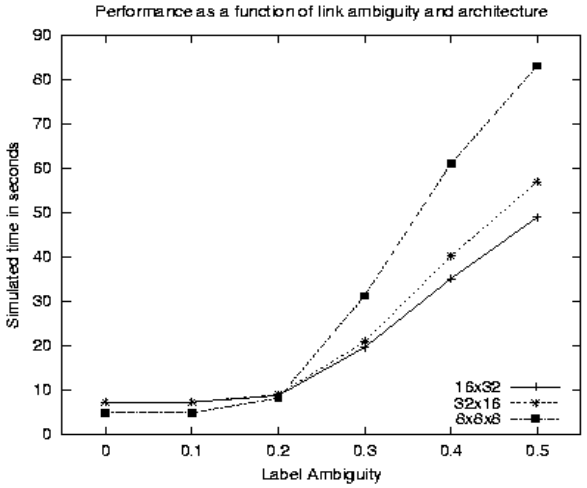


Figure 3 Simulating threshold and opportunistic strategies

Impact of Time Costs

While changing the time costs (250ms for link evaluations and 500ms for link selection and returning to the previous page) will affect absolute simulation times, it is less clear if different time costs will change which architecture produces the fastest times. For example, one may wonder if the 8x8x8 architectures would still produce the slowest times if the link selection cost were double, which may occur for a slower internet connection.

To explore the impact of time costs, we look at the number of link evaluations, link selections and page returns. If independent counts of these actions correlate with the aggregate simulation time, we conclude that varying the time costs have minimal impact on the relative performance of the different architectures. For example, if the 8x8x8 requires more evaluations, more selections and more returns than the other architectures, we know that 8x8x8 will produce slower search times regardless of the time costs.

Looking at the number of evaluations, selections and returns, we see that the 8x8x8 architecture required more of each action (173, 17, and 19 respectively) at the 0.4 noise level than the 16x32 (125, 3, and 5) and the 32x16 (134, 6, and 8). Further experimentation reveals that this relationship holds across all but the lowest noise levels (0.2 and less). We conclude that changing the time costs have no effect on the relative comparisons provided that the noise level is at least 0.3.

Impact of Memory Capacity

Recall that the opportunistic strategy requires the model to store and retrieve threshold values so that the previous threshold can be reinstated upon returning to a parent page. So far, our simulations have assumed that only one threshold value can be restored. Thus, if the model returned to the top level of a three-tier architecture, it would no longer be able to recall the previous threshold and would simply leave the threshold at its current state.

Because this limited memory capacity only hinders performance in a three-tiered site (e.g. 8x8x8), we ran simulations where the memory capacity could hold the additional threshold value so that the previous value could be reinstated when navigating through a three-tiered site. Figure 4 shows the results using the same scale as Figure 3. While we see that the extra memory capacity improves the performance of the 8x8x8 architecture, its navigation is still slower than the two-tiered architectures.

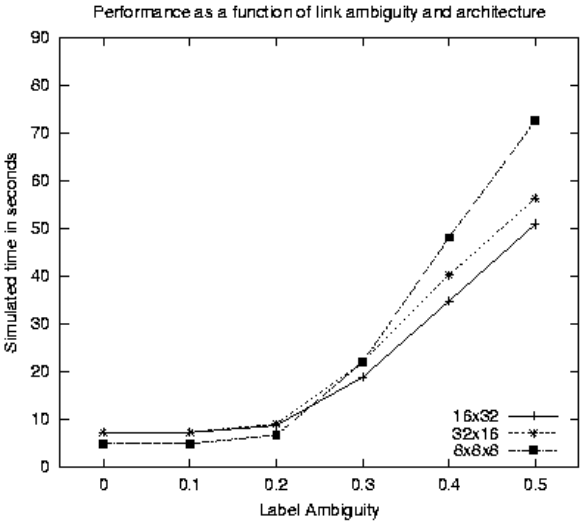


Figure 4 Results using a larger memory capacity

Discussion

We have shown that a simple model of a Web user can provide an excellent account of user behavior and reveal important factors underlying Web usage. The model suggests that link ambiguity interacts with the depth of information architecture to determine site navigation time. As link ambiguity decreases, better performance is found from architectures with deep structures that minimize the number of links searched. As link ambiguity increases, the model shows performance degradations for architectures with deeper structures. The same pattern is characteristic of human users. However, the preference for shallow hierarchies is observed only with sufficient ambiguity in the link

labels and with no ambiguity at the bottom level. Thus, the results of Larson and Czerwinski (1998) may not generalize to large numbers of real Web pages with ambiguity at all levels.

As for Web search strategies, combining threshold-based selection with opportunistic search strategies produced simulated times that are very close to observed times for 0.4 noise level. This also corresponds to the behavior of several users searching a department store site in the pilot study mentioned above. We recognize the need for converging methods to independently determine link ambiguity and are exploring theoretical and empirical methods of estimating actual values.

To make time predictions, our model assumes plausible time costs for link evaluation, link selection and returning to the previous page. By noting the actual counts for these operations, our simulations help us understand what happens when the link selection time is significantly longer, as would be the case for a slow internet connection. We found, however, that the time costs have no effect on the relative comparisons provided that the noise factor is at least 0.3. This suggests that a slower internet connection does not impact the relative advantage of shallow architectures when significant link ambiguity is present, at least for the case where no noise is present at the bottom level.

Our simulations also aid our understanding of how human memory impacts effective navigation. Increasing the model's memory capacity improved performance for the deep (8x8x8) structure but left the other two architectures largely unaffected. This suggests that memory is more useful in keeping track of site architecture than in searching within a page. Since searching a page is facilitated by visual cues (e.g., changes in the color of previously selected links) users can avoid reliance on memory. Visual cues are typically not present to remind users of the names and locations of previous links. The interaction of structure with memory capacity indicates further that simple heuristics for representing capacity are insufficient to capture memory phenomena of importance. Instead, it is necessary to examine how the structure of information sites provides aids to memory. Our analysis contrasts with previous advice suggesting that the number of links per page should be limited to 10 (Rosenfeld & Morville, 1998) (see Larson & Czerwinski, 1998, for a discussion based on experimental results).

We have shown that a simple model of a user interacting with a simplified Web site can reveal important factors that affect usability and can support the investigation of the interactions between those factors across a wide range of conditions. What we have presented is not a comprehensive model of Web navigation. No attempt is made to account for how people scan a page, or evaluate link labels or images. By abstracting these processes, and representing only their functionality, the model focuses on understanding how information architecture affects the navigation process. As an approximation of user navigation, the model can account for a range of human behaviors by varying likelihood factors in its site representations. We have shown that the model provides a good approximation of the behavior of the common (modal) user. By varying parameters it should be possible to extend the model to account for alternate strategies.

References

Ballard, D. H., Heyhoe. M. M., Pook, P. K., & Rao, R. P. N. (1997). Diectic codes for the embodiment of cognition. Behavioral and Brain Sciences, 20, 723-767.

Byrne, M. D., John, B. E., Wehrle, N. S., & Crow, D. C. (1999). The tangled web we wove: A taskonomy of WWW use. Proceedings of CHI'99 Human Factors in Computing Systems (pp. 183-190). New York: ACM press.

Card, S. K., Moran, T. P., & Newell, A. (1983). The Psychology of Human-Computer Interaction. Hillsdale, NJ: Lawrence Erlbaum.

Larson, K. & Czerwinski, M. (1998). Web page design: Implications of memory, structure, and scent for information retrieval. Proceedings of CHI'98 Human Factors in Computing Systems (pp. 25-32). New York: ACM press.

Lynch, G., Palmiter, S., & Tilt, C. (1999) The max model: A standard Web site user model. Proceedings of the 5th Annual Human Factors and the Web Conference. Retrieved February 5, 2001 from the World Wide Web: http://zing.ncsl.nist.gov/hfweb/proceedings/lynch.

Miller, C. S. & Remington, R. W. (2000a). A computational model of Web navigation: Exploring interactions between hierarchical depth and link ambiguity. Proceedings of the 6th Conference on Human Factors and the Web. Retrieved February 5, 2001 from the World Wide Web: http://www.trisbc.com/hfweb/miller/article.html.

Miller, C. S. & Remington, R. W. (2000b). Exploring Information Navigation Strategies with a Computational Model (Report No. 00-004). Chicago: DePaul University, School of Computer Science, Telecommunications, and Information Systems.

Miller, D. P. (1981). The depth/breadth tradeoff in hierarchical computer menus. Proceedings of the Human Factors Society (pp. 296-300).

Peck. V. A. & John, B. E. (1992). Browser Soar: A computational model of a highly interactive task. Proceedings of CHI'92 Human Factors in Computing Systems (pp. 165-172). New York: ACM press.

Rosenfeld, R. & Morville, P. (1998). Information Architecture for the World Wide Web. Sebastopol, CA: O'Reilly & Associates.

Emergence of Effects of Collaboration in a Simple Discovery Task

Kazuhisa Miwa (miwa@cog.human.nagoya-u.ac.jp)
Graduate School of Human Informatics, Nagoya University
Nagoya, 464-8601 JAPAN

Abstract

We discuss the effects of collaboratively finding a target in a simple discovery task, using the Wason's 2-4-6 task. We control the following two factors: hypothesis testing strategies that participants use, and the nature of targets that they find. First, we propose, through computer simulations, a hypothesis on a situation in which the effects emerge. Then we verify the hypothesis by psychological experiments. Last, we generalize, through theoretical analysis, the findings obtained by the two empirical approaches above. As the result, it has been concluded that the effects of collaboration emerge in the following situations: (1) two participants repeatedly conduct a positive test for finding a general target, and (2) each of them maintains a different hypothesis.

Introduction

In psychological studies on scientific discovery, relatively simple tasks, such as the Wason's 2-4-6 task and New Elusis, have been so far used (Gorman, 1992; Newstead & Evans, 1995). In recent days, using those tasks, the effects of collaboration have been empirically discussed when several participants collaboratively find a target.

We usually think that working together provides positive effects. However, those empirical results obtained in the psychological studies above do not necessarily support the intuitive prediction.

In these studies, the performances (the proportions of correct findings) in the single condition in which a single subject performs the task and those in the collaborative condition in which a group of n subjects collaboratively performs the task are compared. In this comparison, even when the latter performance exceeds the former, the advantage may be introduced not by the interaction among the subjects, but simply by the quantity of the subjects. That is, in the latter case n solutions (final hypotheses) by the n subjects are considered, and the probability of that at least one of them is accidentally identical to the target is much higher than that in the former case. So we should consider the independent condition in which n participants independently perform the task without interaction. The performance in the independent condition can be theoretically calculated from the performance in the single condition. That is, the probability of that at least one of n subjects reaches the solution is $1 - (1 - p)^n$ where the probability of each subject's finding the correct target is p ($0 < p < 1$). We utilize this score as the performance in the independent condition.

Table 1 reviews the comparison of the performances in the single, independent, and collaborative conditions in the preceding studies (Freedman, 1992; Laughlin & Futoran, 1985; Laughlin & McGlynn, 1986; Laughlin, VanderStoep, & Hollingshead, 1991; Laughlin, Bonner, & Altermatt, 1998; Okada & Simon, 1997). Table 1 shows that the performance in the collaborative condition cannot exceed that in the independent condition in almost all cases.

In this study, we will discuss states in which the effects of collaboration emerge based on the results above. As an approach, first we will propose a hypothesis on when the effects of collaboration appear by computer simulations using a computational model that solves the Wason's 2-4-6 task (Wason, 1960). Then we will verify the hypothesis by psychological experiments. Last we will generalize the empirical findings by theoretical task analysis, and discuss why the effect emerges only in the specific situation.

Fundamental issues

Klayman & Ha, in their paper in 1987, gave some decisive answers to several historical questions that had been discussed in the psychological studies using traditional discovery tasks such as the Wason's task (Klayman & Ha, 1987). One of their major conclusions was that there was substantial interaction between the nature of found targets and the effectiveness of hypothesis testing strategies used by subjects. So in this study we will control these two factors in the following experiments.

First, we briefly explain some important concepts about the two factors: the nature of targets that subjects should find and the hypothesis testing strategies that subjects use.

The nature of targets: we categorize targets used in our experiments from the viewpoint of their generality. General targets are defined as the targets, the proportion of whose member (positive instances) to whole instances in the searched space is large. On the other hand, specific targets are defined as the targets, the proportion of whose member is small. For example, the proportion of the instances fitted to "the product is even" and "three evens" to all possible

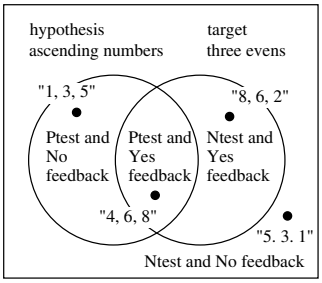
Table 1 Comparisons of the performances in the single, independent, and collaborative conditions in the preceding studies.

	Laughlin (1985)	Laughlin (1986)	Laughlin (1991)					Laughlin (1998)	Freedman (1992)		Okada (1997)
# of group members	4	4	4					4	4		2
task	New Elusis	New Elusis	New Elusis					New Elusis	2-4-6 task		simulated molecular genetics laboratory
single	0.15	0.19	0.06	0.13	0.15	0.14	0.16	0.16	0.33	0.08	1.7
independent	0.47	0.57	0.2	0.3	0.38	0.3	0.41	0.38	0.80	0.28	2.1
collaborative	0.35	0.34	0	0.1	0.2	0.1	0.41	0.34	0.83	0.67	2.9

instances are 7/8 and 1/8 respectively. So the former is an example of a general target and the latter is an example of a specific target.

Hypothesis testing strategies: There are two types of hypothesis testing: a positive test and a negative test. The positive test (Ptest) is conducted by an instance subjects expect to be a target. That is, Ptest is hypothesis testing using a positive instance for the hypothesis; the negative test (Ntest) is hypothesis testing using a negative instance. For example, when a hypothesis is "ascending numbers", hypothesis testing, using an instance, "1, 3, 9", is Ptest; hypothesis testing, using "1, 5, 2", is Ntest.

In the following description, for avoiding the confusion of the basic concepts, we define *Yes* and *No* instances as members and non-members (positive and negative instances) for targets that subjects do not know. On the other hand, we also define *Positive* and *Negative* instances as members and non-members for hypotheses that subjects form. When a subject conducts an experiment using a *Positive* instance for his/her hypothesis, and knows, through the feedback from an experimenter, that the instance is a *Yes* instance for a target, we call that the subject receives *Yes* feedback as a result of his/her *Ptest.*



hypothesis	target: Yes	target: No
Ptest	confirmation	disconfirmation
Ntest	disconfirmation	confirmation

Figure 1 Patterns of confirmation and disconfirmation.

Klayman et. al. summarized the states when a subject's hypothesis was disconfirmed. Figure 1 illustrates those states in the example situation of that the target is "three evens" and the subject's hypothesis is "ascending numbers". When the subject conducts Ptest, using an instance, "1, 3, 5", and receives No feedback, his/her hypothesis is disconfirmed. Another state of disconfirmation is introduced by the combination of Ntest and Yes feedback, using an instance, "8, 6, 2". On the other hand, the states of confirmation are introduced by the combination of Ptest and Yes feedback, using "4, 6, 8", and the combination of Ntest and No feedback, using "5, 3, 1".

Computer simulations

First, we will propose a hypothesis on when advantages of collaboration appear by computer simulations.

Specifications of the model

In the following, we will explain only the summary of our model. Detailed specifications of the model can be seen in our other papers (Miwa, 1999).

The model was constructed on the interactive production system architecture that we had developed for simulating collaborative problem solving processes. The architecture primarily consists of five parts; production sets of System A, production sets of System B, the working memory of System A, the working memory of System B, and a common shared blackboard. Two systems interact through the common blackboard. That is, each system writes elements of its working memory on the blackboard and the other system can read them from the blackboard.

The model has knowledge on the regularities of three numerals. The knowledge is organized as the dimension-value lists. For example, "continuous

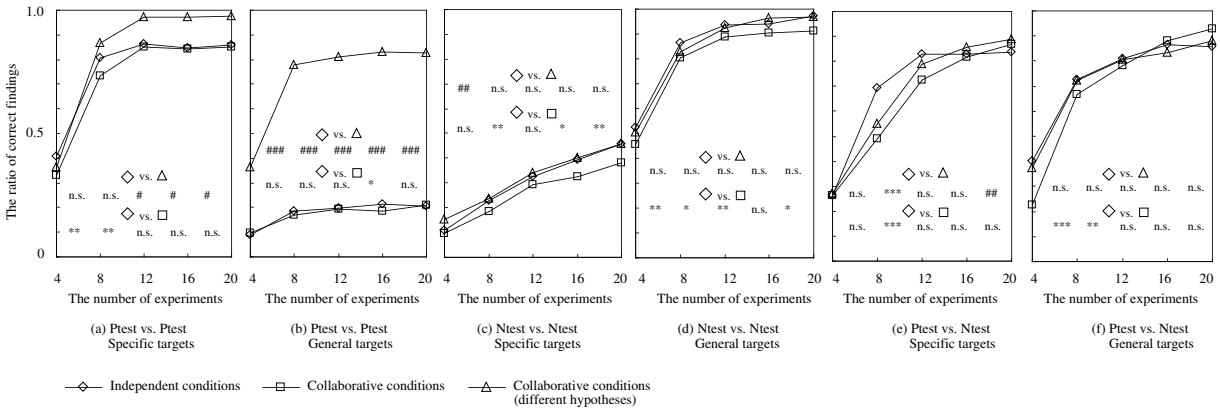


Figure 2 Results of the computer simulations.

evens", "three evens", and "the first numeral is even" are example values of a dimension, "Even-Odd". The dimensions the model uses are: Even-Odd, Order, Interval, Range of digits, Certain digit, Mathematical relationship, Multiples, Divisors, Sum, Product, and Different.

Basically the model searches the hypothesis space randomly in order to generate hypotheses. However, three hypotheses, "three continuous evens", "the interval is 2", and "three evens" are particular. Human subjects tend to generate these hypotheses at first when the initial instance, "2, 4, 6", is presented. So our model also generates these hypotheses first prior to other possible hypotheses.

As for a way of hypothesis verification, the principle on when a model's hypothesis is disconfirmed and a next hypothesis is reconstructed is based on the Klayman & Ha's schema shown in the previous section.

The design of simulations

In our computer simulations, we let the two systems find 35 kinds of targets. Examples of the targets are: three continuous evens, ascending numbers, the interval is 2, single digits, the second numeral is 4, first numeral times second numeral minus 2 equals third numeral, multiples of 2, divisors of 24, the sum is a multiple of 12, the product is 48, and three different numbers. The initial instance was "2, 4, 6". For each target, we executed 30 simulations to calculate the percentage of correct solutions.

The computer simulations were conducted based on the following 2 * 3 experimental design.

The nature of targets: We divided the 35 targets into two categories: (a) 17 specific targets and (b) 18 general targets.

Hypothesis testing strategies: Three combinations of hypothesis testing strategies were investigated. They were (a) Ptest and Ptest, (b) Ntest and Ntest, and (c) Ptest and Ntest.

Results of the simulations

Figure 2 shows the results of the computer simulations. The horizontal axis of the figure indicates the number of experiments, that is, the number of generated instances whereas the vertical axis indicates the proportion of correctly finding the 17 specific targets and the 18 general targets.

In Figure 2, the performances in the independent condition and those in the collaborative condition are compared. In the independent condition, we regard that the targets are correctly found when at least one of the two systems, each of which independently tries to find the targets without interaction, reaches the correct solution.

In the collaborative condition, experiments are alternately conducted. Through each simulation, one system generates the half of whole instances; and the other half is generated by the other system. Each experimental result is shared by both two systems, that is, each system knows whole generated instances with Yes or No feedback that is given to each instance.

In addition, the collaborative condition is subdivided into the following two sub-conditions. In one sub-condition, each system simply alternately conducts experiments, not referring to another hypothesis that the other system forms. In this sub-condition, two systems share only the experimental space. In the other sub-condition, one system tries to form a different hypothesis, referring to another hypothesis of the other system. In the latter sub-condition, two systems share the hypothesis space in addition to the experimental space (Klahr &Dunber, 1988).

In the figure, the results of statistical analysis are also indicated. The upper row indicates the difference

between the performances in the independent condition and those in the collaborative condition where two systems try to form different hypotheses, whereas the lower row indicates the difference between the performances in the independent condition and those in the collaborative condition where each system does not refer another hypothesis of the other system. The asterisks show the advantage of the independent condition whereas the sharps show the advantage of the collaborative condition. Three levels of significance are used: ### (or ***) for $p < .01$, ## (or **) for $p < .05$, and # (or *) for $< .1$. No significance is indicated with n.s.

Figure 2 indicates that the performance in the collaborative condition exceeds that in the independent condition only when (1) both systems use the Ptest strategy for finding general targets, and (2) both systems try to form different hypotheses, sharing their hypothesis space. In the other cases, the effect of collaboration is not remarkable.

Psychological experiments

To verify the results of the computer simulations in the previous section, we conducted psychological experiments.

Design and procedure

A total of 136 subjects participated in the experiments. Each of them was assigned to each of the following five experimental conditions: (1) the single Ptest condition where a single participant solved a task using Ptest, (2) the single Ntest condition, (3) the collaborative Ptest and Ptest condition where two participants, both of whom were required to use Ptest, collaboratively solved a task, (4) the collaborative Ntest and Ntest condition, and (5) the collaborative Ptest and Ntest condition. Each subject solved two problems. In one problem, "three evens", as a specific target, was discovered. In the other problem, "three different numbers", as a general target, was discovered. The order of the problems was counter-balanced. Twenty-four trials (experiments) were permitted for finding each target. The experimental design is summarized in Table 2.

Table 2 The number of subjects and pairs in each experimental condition.

	single		pair		
	Ptest	Ntest	Ptest v.s. Ptest	Ntest v.s. Ntest	Ptest v.s. Ntest
specific	17(15)	18(14)	16(15)	15(11)	17(11)
general	17(10)	17(12)	17(12)	16(9)	17(9)

In the following discussion, we exclude the results of the subjects who did not follow the experimental instruction requiring to use each hypothesis testing strategy. Table 2 shows the number of subjects (or pairs) assigned to each experimental condition, and, in parenthesis, the number of them who correctly follow the Ptest and Ntest instruction.

Results of the experiments

Figure 3 indicates the experimental results, using the same format of Figure 2. In Figure 3, experimental results that were actually obtained in the experiments are the performances in the collaborative condition (collaborative conditions in (a) through (f)), and those in the single condition where both participants used the same hypothesis testing strategy (single conditions in (a) through (d)). On the other hand, the performances in the single condition, where each subject used a different strategy, are the average scores of the performances in the single Ntest condition and those in the single Ptest condition (single conditions in (e) and

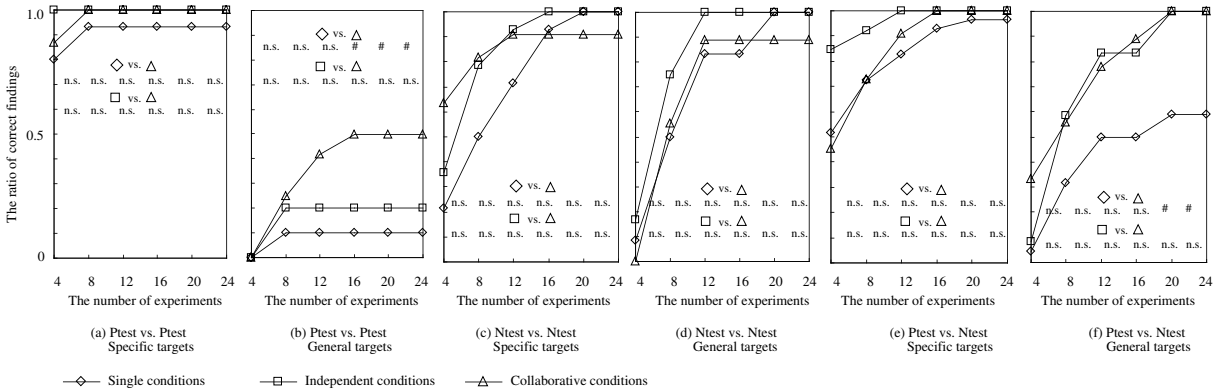


Figure 3 Results of the psychological experiments.

(f)). Additionally the performances in the independent condition are calculated from the performances in the single condition by using the similar procedure indicated in the introduction (independent conditions in (a) and (f)).

As for statistical analysis, the upper row indicates the difference between the performances in the single condition and those in the collaborative condition, whereas the lower row indicates the difference between the performances in the independent condition and those in the collaborative condition.

The statistical analysis shows, in every combination of the hypothesis testing strategies, that the performances in the collaborative condition cannot exceed those in the independent condition. However, only in the combination of Ptest and Ptest for finding the general target, the performance in the collaborative condition exceeds that in the single condition, and a tendency of the advantage of the collaborative condition over the independent condition is observed even though the statistical analysis does not indicate the significant difference.

Next, to confirm the effect of two subjects' forming different hypotheses, we will conduct the following additional analysis. First, we divide the subjects in each collaborative condition into two groups: the subjects who found the correct target earlier and those who did later. The latter group includes those who did not find correct target. Then we measure the average of the proportion of that the subjects in each group maintained different hypotheses through the trials till reaching the solution. Figure 4 shows the result. What we note is that the effect of forming different hypotheses appears in the combination of Ptest and Ptest, especially when finding the general target, whereas this effect does not appear in the combination of Ntest and Ntest. These results are consistent with the findings of the computer simulations.

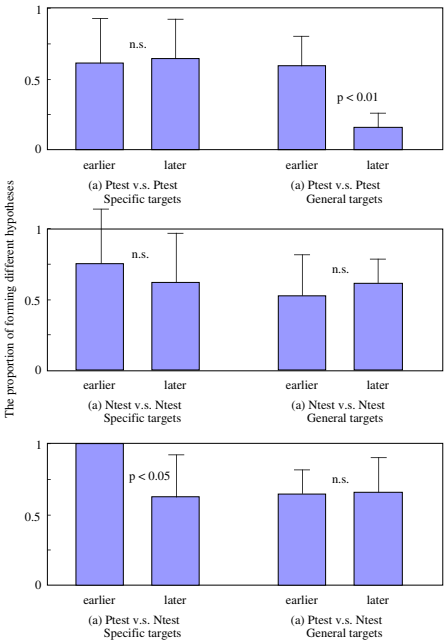


Figure 4 Proportion of forming different hypotheses in the earlier finding group and the later finding group.

Theoretical analysis

Why does the advantage of collaboration emerge only when both participants, for finding the general targets, repeatedly conduct Ptest? We will discuss the reason based on the Klayman & Ha's framework of analysis.

Klayman et. al. indicated, by their mathematical analysis, that Ptest was an effective heuristic for finding specific targets; on the other hand, Ntest was effective for finding general targets.

When a target is general, the possibility of receiving Yes feedback is high in the experiments. In the situation, it is difficult that Ptest introduces disconfirmation because the combination of Ptest and Yes feedback introduces confirmation. So Ptest often prevents subjects from finding general targets. In addition, Ntest is an ineffective strategy for finding specific targets because subjects very often receive No feedback as a result of their Ntest. The collaboration of two systems can compensate for these week points of hypothesis testing strategies.

Let us consider the collaborative condition in which both two systems (or two subjects), System A and System B, alternately conduct Ptest, and the systems have different hypotheses. In this situation, it happens that a positive instance for a hypothesis of System A, HA, corresponds to a negative instance for another hypothesis of System B, HB. For example, when a hypothesis HA is "the interval is 2" and a hypothesis HB is "ascending numbers", an instance, "2,

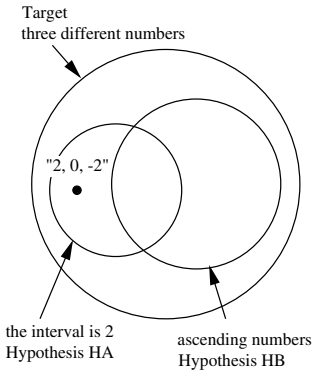


Figure 5 A situation of that Ptest of one system, System A, introduces Ntest of the other system, System B.

0, -2", is this kind of instance (see Figure 5).

When System A conducts Ptest, using this instance, it happens that for System B Ntest is introduced by the instance generated by System A. As a result, Yes feedback introduces disconfirmation of the hypothesis HB because the combination of Ntest for HB and Yes feedback is carried. This brings the effect of collaboration when two systems, both of which use Ptest, find general targets.

An important point is that this function emerges in the interaction between two systems. This advantage is not introduced as the effect of the quantity of the systems. That is, the advantage is not the effect of that the number of systems in the collaborative condition is twice as many as that in the single condition. As you can confirm in Figure 5, when each system independently conducts Ptest, a hypothesis of each system is never disconfirmed. Chances of hypothesis disconfirmation can be introduced only through the collaboration of two systems.

A next question is why this kind of effect does not appear in the combination of Ntest and Ntest when finding specific targets where the probability of subjects' receiving No feedback is very high.

If the above-mentioned type of interaction between two systems emerges in the combination of Ntest and Ntest, the situation in which Ntest of System A introduces Ptest for System B should happen. However, generally speaking, members (positive instances) of a hypothesis is much fewer than the non-members (negative instances). So the possibility of constructing the situation in which Ntest of one system accidentally introduces Ptest for the other system, where the effect of the Ntest and Ntest collaboration appears, is much lower than the possibility of constructing the situation in which Ptest of one system introduces Ntest for the other system, where the effect of the Ptest and Ptest collaboration appears. This is the reason why only the combination of Ptest and Ptest introduces the effect of collaboration.

Conclusions

In the introduction of this paper, we indicated that the effects of collaboration rarely appeared in the psychological experiments, using orthodox simple discovery tasks. We empirically demonstrated a situation in which those effects of collaboration emerged, and theoretically discussed why the effects were introduced. Concretely, we indicated that the effects appeared when both subjects (systems) verified their hypotheses by using Ptest for finding general targets. This result is more interesting, as a finding on collaborative discovery, when we note that humans have a cognitive bias of tending to use Ptest more frequently.

Our empirical findings and theoretical discussions conclude that (1) generally speaking, simply solving a problem together rarely introduces the effects of collaboration, (2) to introduce the effects of collaboration, it is needed that the interaction between collaborative systems brings new abilities, such as a function for introducing disconfirmation of their hypotheses, which are not involved in each individual system, and (3) the possibility of bringing those abilities depends on natures of objects that systems investigate, strategies and heuristics that systems use, and the relation between these factors.

References

Freedman, E. (1992). Scientific Induction: Individual versus Group Processes and Multiple Hypotheses. Proceedings of the 14th annual meeting of cognitive science society, 183-188.

Gorman, M. (1992). Simulating science: heuristics, mental models, and technoscientific thinking. Indiana university press.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. Cognitive Science, 12, 1-48.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. Psychological Review, 94, 211-228.

Laughlin, P.R., & Futoran, G.C. (1985). Collective induction: Social combination and sequential transition. Journal of Personality and Social Psychology, 48, 608-613.

Laughlin, P. R., & McGlynn, R. P. (1986). Collective induction: mutual group and individual influence by exchange of hypotheses and evidence. Journal of Experimental Social Psychology, 22, 567-589.

Laughlin, P. R., VanderStoep, S. W., & Hollingshead, A. B. (1991). Collective versus individual induction: recognition of truth, rejection of error, and collective information processing. Journal of Personality and Social Psychology, 61(1), 50-67.

Laughlin, P. R., Bonner, B. L., & Altermatt, T. W. (1998). Collective versus individual induction with single versus multiple hypotheses. Journal of Personality and Social Psychology, 75 (6), 1481-1489.

Miwa, K. (1999). Collaborative Hypothesis Testing Process by Interacting Production Systems, Lecture Notes of Artificial Intelligence, 1721, 56-67.

Newstead, S., & Evans, J. (Eds.). (1995). Perspectives on Thinking and Reasoning. UK: Lawrence Erlbaum Associates Ltd.

Okada, T., & Simon, H. (1997). Collaborative discovery in a scientific domain. Cognitive Science, 21, 109-146.

Wason, P. (1960). On the failure to eliminate hypotheses in a conceptual task. Quarterly journal of experimental psychology, 12, 129-140.

Effects of Competing Speech on Sentence-Word Priming: Semantic, Perceptual, and Attentional Factors

Katherine Moll (**Katherine.Moll@keble.ox.ac.uk**)
Keble College, University of Oxford
Oxford OX1 3PG United Kingdom

Eileen Cardillo (**Eileen.Cardillo@st-johns.ox.ac.uk**)
Department of Experimental Psychology, University of Oxford
South Parks Road, Oxford OX1 3UD United Kingdom

Jennifer Aydelott Utman (**Jennifer.Utman@psy.ox.ac.uk**)
Department of Experimental Psychology, University of Oxford
South Parks Road, Oxford OX1 3UD United Kingdom

Abstract

This study examined the effects of a competing signal on sentence-word priming using an auditory lexical decision paradigm. Previous studies have suggested that the facilitatory component of the sentential priming effect is particularly sensitive to acoustic distortions that reduce the perceptibility of the sentence context, whereas the inhibitory component is more sensitive to increased attentional demand. Three competing signal conditions were compared: forward speech presented to a different ear, backward speech presented to a different ear, and forward speech presented to the same ear. The results demonstrate that the competing signal has different effects on priming depending upon the semantic content of the signal and its perceptual isolability from the sentence context.

Introduction

The influence of sentence meaning on lexical processing has been studied extenstively in sentence-word priming studies, which have demonstrated that sentence context has both facilitatory and inhibitory effects on responses to single words (e.g., Stanovich & West, 1983; Meyer & Schvaneveldt, 1976; Simpson, Peterson, Casteel, & Burgess, 1989; Duffy, Morris, & Henderson, 1989). In other words, reaction times for words preceded by a highly congruent sentence context are faster than those preceded by a neutral context, whereas responses to words presented in an incongruent are slowed relative to a neutral context. Further, facilitatory and inhibitory effects tend to emerge under different experimental conditions in both sentence-word and word-word priming studies. Specifically, facilitation effects emerge at brief stimulus onset asynchronies (SOAs), and are generally insensitive to expectancy and processing strategies (Neely, 1991; Utman & Bates, 1998a,b), to the extent that facilitation is observed even when a prime stimulus is presented briefly and then masked, such that subjects are not consciously aware of the identity of the prime (Neely, 1991). In contrast, inhibitory effects are more pronounced at longer SOAs, and are sensitive to factors that affect expectancy, such as the proportion of related/unrelated test trials in the stimulus set (Neely, 1991), and tend to be reduced in populations with limited attentional capacity (e.g., Faust, Balota, Duchek, Gernsbacher, & Smith, 1997). Based on this evidence, researchers have argued that facilitation occurs rapidly and requires little in the way of attentional and/or processing resources, whereas inhibition occurs later in the time course of lexical access, and may reflect more controlled or strategic processing (Faust & Gernsbacher, 1996; Gernsbacher, 1997; Neely, 1991; Utman & Bates, 1998a,b).

Recent studies have revealed that the facilitatory and inhibitory components of the sentence-word priming effect respond differently to acoustic distortion of the semantic context. Specifically, facilitatory effects are particularly sensitive to distortions that reduce the intelligibility of the acoustic signal, whereas inhibitory effects are more sensitive to distortions that reduce processing time and/or increase attentional demand (Utman & Bates, 1998 a,b; Utman, Dick, Prat, & Mills, 1999). For example, low-pass filtering of the sentence context significantly reduces facilitation of congruent targets (Utman & Bates, 1998a,b). In contrast, temporal compression of the context significantly reduces inhibition of incongruent targets, but has no effect on facilitation (Utman & Bates, 1998a,b). These findings suggest that facilitation effects are more dependent on bottom-up information from the acoustic signal, whereas inhibition effects are more dependent on attentional resources.

In the present study, we sought to investigate how competing speech influences the facilitatory and inhibitory components of the sentence-word priming

effect. The separate influences of perceptual and attentional factors on facilitation and inhibition have particular implications for competing speech. It is generally recognised that competing speech places an increased demand on processing resources, as the listener must selectively attend to one signal while suppressing another. For example, Downs & Crum (1978) found that competing speech significantly increased effort required for auditory learning, although actual performance was not affected. They concluded that competing speech increased attentional load, thereby decreasing resources for speech processing. More recently, Connolly, Phillips, Stewart, & Brake, (1992) found that hearing competing speakers resulted in a decreased ability to process semantic features of the target speech, indicating that the increased attentional demand of competing speech directly affects semantic processing.

Competing speech may also be relatively more demanding than a competing signal with no semantic content, because the speech signal will activate linguistic representations that are in conflict with the attended signal. Unattended speech has been found to be processed at a semantic level by Moray (1959) as well as Bentin, Kutas, & Hillyard (1995). Further, Garstecki & Mulac (1974) found that subjects had more difficulty on an auditory discrimination task when competing speech was played forwards than when it was played backwards, although subjects described the backwards speech as sounding like language. The interference effect was attributed to the semantic content of the forward speech. In addition, Harris, Benedict, & Leek (1990) observed that performance on a language-based task was more severely disrupted by more intelligible competing speech, and that performance was worst with only one competing speaker, as it is easier to extract semantic information from the competing signal of a single voice than several.

In addition to placing increased demands on attentional resources, competing speech may also affect the perceptibility of the attended signal by masking the spectral frequencies of the signal, thereby interfering with the bottom-up encoding of phonetic contrasts. This effect may be particularly pronounced when the two competing signals come from a similar spatial location, making the target signal more difficult to isolate.

These possibilities were explored in a sentence-word priming study, in which the sentence context was presented in isolation or in the presence of a competing signal. Lexical decision responses to target words were compared when the targets were presented in congruent, neutral, and incongruent sentence contexts. Given the framework described above, it was possible to predict the effect of different forms of auditory competition on lexical decision following biasing sentence contexts.

When the context was easily isolable from the competing signal, we expected that the primary effect of competing speech would be to increase attentional demand, leaving the perceptibility of the context relatively unaffected. As a result, we predicted that strategic or attentional inhibitory processes would be severely compromised, while facilitation would remain intact. This hypothesis was tested by presenting the semantic context and the competing speech signal to different ears. In this case, we predicted no significant change in reaction times to contextually congruent targets, but a significant decrease in inhibition for contextually incongruent words.

We expected this release from inhibition to be more pronounced for a competing signal with semantic content, as a meaningful signal should place relatively greater demands on available resources for semantic processing. To confirm this prediction, we compared the effect of competing speech with that of a similar competing signal with no semantic content, i.e. backward speech. Since backward speech does not appear to greatly tax attentional resources despite its speech-like properties, we predicted no significant difference in reaction times between targets preceded by sentence contexts presented in isolation and targets preceded by sentence contexts accompanied by competing backward speech in a different ear.

We also considered the effect of competing speech when presented in the *same* auditory channel as the sentence contexts, by presenting forward speech in the same ear as the semantic context. In this case, we expected that the competing signal would have a masking effect, disrupting the peripheral encoding of the speech input, as well as making it more difficult to attend only to the appropriate sentence. Consequently, we predicted that both facilitatory and inhibitory effects of sentence context would be affected.

Method

Participants

Thirty-six undergraduates (15 male and 21 female) at Oxford University participated in the experiment. All participants were right-handed, native British English speakers. None reported any hearing impairments.

Stimuli

The stimuli consisted of 30 target words, 30 nonwords, a neutral sentence, and 60 highly constraining sentence contexts (30 to be paired with targets and 30 with nonwords).

The targets were one-syllable words containing 3-5 phonemes (mean = 3.27, SD = 0.64) with a mean Kucera-Francis print frequency of 139 (SD = 99) (Kucera and Francis, 1967), a mean London-Lund spoken frequency of 14 (SD = 81) (Brown, 1984), and a mean concreteness rating of 546 (SD = 81) as specified in the MRC Psycholinguistic Database (Coltheart, 1981). To avoid possible morphological and morpho-phonological constraints of determiners (*a/an*, *the*), mass nouns (e.g., *blood*, *dust*) were excluded, and all targets were consonant-initial. The nonword distracter targets consisted of phonologically permissible one-syllable nonsense items, which did not differ significantly from the targets in terms of number of phonemes (mean = 3.33, SD = 0.61).

The sentence contexts matched with the word targets were approximately ten syllables in duration (mean = 9.47, SD = 2.66), containing a maximum of six content words (mean = 3.37, SD = 1.07), and a maximum of three words related to the congruent target (mean = 1.13, SD = 0.51). There was no significant difference in length or number of content words between sentences paired with word targets and sentences paired with nonword distractors.

In the congruent condition, word targets were matched with the appropriate sentence context (e.g., *He wanted to come in, but she refused to open the...* - DOOR). In the incongruent condition, word targets were matched with a sentence context appropriate to another target in the stimulus set (e.g., *He wanted to come in, but she refused to open the...* - HORSE). Pilot analyses revealed that word targets had a mean cloze probability of 100% (SD = 0%) when presented in the correct context. Therefore, if a target was presented in an incongruent context its cloze probability was 0%. A neutral sentence context, providing no semantic cues with regard to the target (*The next item is...*), was created to serve as a neutral baseline.

Each participant received one of six randomised lists, each containing 60 trials. Across lists, each target appeared in each condition (congruent alone, congruent + competing signal, neutral alone, neutral + competing signal, incongruent alone, and incongruent + competing signal), with no target or biasing context appearing more than once per list. In each of the three semantic context conditions, half of the sentences were presented without a competing signal (presented alone in only one ear), or with a competing signal (presented in one ear with the competing signal presented in the same ear or the other ear). Thus, semantic bias (congruent, neutral, or incongruent) and competing signal (present or absent) served as within-subjects variables (each participant received all six conditions). This pattern also applied to the thirty nonwords and their sentence contexts in each list.

The stimuli were produced by native speakers of British English. To distinguish each target clearly from the preceding context, the words and nonwords were produced by a male speaker, and the sentences were produced by a female speaker. The stimuli were recorded onto digital audio tape in an Industrial Acoustics 403-A audiometric chamber with a Tascam DA-P1 Digital Audio Tape recorder and a Sennheiser ME65/K6 supercardioid microphone and pre-amp at gain levels between -6 and -12 db. The recorded stimuli were then digitised via digital-to-digital sampling onto a Macintosh G4 computer via a Digidesign audio card using ProTools LE software at a sampling rate of 44.1 kHz with a 16-bit quantization. The waveform of each sentence, target, and nonword was then edited and saved in its own mono-audio file. All the stimulus files were converted into 16-bit 22.05 kHz stereo files in SoundEdit16 and saved in System 7 format.

Competing Speech Conditions

A passage from the book *Profit Patterns* (Slywotzky, 1999) was also recorded and edited on the same equipment and under the same conditions (including gain levels) but using a different female speaker. This recording was then used for both the forward and backward competing speech conditions.

In the forward competing speech condition (different ear), copies of all the stereo sentence files were made and segments of competing speech of the same duration as the sentence context were excised at random and inserted on the blank track in the stereo sound file.

In the backward competing speech condition (different ear), the same competing speech was used as in the forward speech condition, played backward. This was achieved using the backwards function in SoundEdit16. This condition was intended to produce auditory interference with the same frequency spectrum as speech but without semantic content.

In the forward competing speech condition (same ear), the sentence contexts were mixed with the forward speech and presented through a single channel. All the original stereo files containing forward competing speech were converted into new stereo files in which the two tracks had been mixed together using the mix function in SoundEdit16.

Type of competing signal (forward/different ear, backward/different ear, and forward/same ear) served as a between-subjects variable (each participant was assigned to one of the three conditions).

Procedure

The test trials were presented auditorily with an inter-stimulus interval of 1500 ms on a Macintosh G4 computer using SuperLab software. The stimuli were

presented through Sennheiser HD 25-1 headphones via a Sirocco VideoLogics amplifier in a sound-protected testing room. Reaction times (RTs) and accuracy were recorded in SuperLab from a Cedrus RB-610 response box. Subjects were instructed to respond as quickly and accurately as possible after hearing the target word, and to press a green button if they heard a real English word or a red button if they heard a nonword. Whether the participant heard the context sentence in the left or right ear was counterbalanced across both lists and male and female subjects.

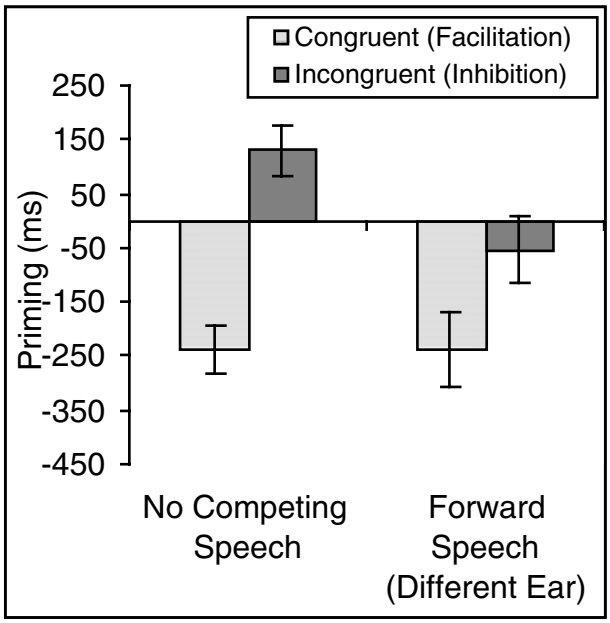


Figure 1. Effect of forward competing speech (different ear). Error bars indicate standard error.

Results

To minimise the effect of outliers in the RT data, the median RT for correct responses was calculated for each subject in each condition for use in the statistical analyses (Wilcox, 1992; Ulrich & Miller, 1994). The magnitude of the priming effect was obtained for both RTs and accuracy by subtracting average median values in the biasing conditions from average median values in the neutral baseline condition for each subject. Four-way subject and item analyses of variance (ANOVAs) comparing the magnitude of the priming effect across conditions were conducted for accuracy (percent correct) and RTs (milliseconds), with competing signal (present vs. absent) and semantic bias (congruent vs. incongruent) as within-subjects variables and type of signal (forward/different ear vs. backward/different ear vs. backward/same ear) and context ear (left vs. right) as between-subjects variables. There was no main effect of context ear in either the accuracy or RT analyses. Further, there was no significant interaction of context ear with competing signal, type of signal, or semantic bias, and no four-way interaction, in either the accuracy or RT analyses. Thus, there was no significant right ear advantage for semantic context, and no significant difference between the pattern of performance obtained across conditions when the semantic context was presented to the right ear and the pattern when the semantic context was presented to the left ear.

A significant interaction of competing signal x semantic bias x type of signal emerged (Subject $F(2,30) = 4.69$, $p < .05$) emerged, although this effect failed to generalise across items. The accuracy analyses revealed no significant three-way interaction, nor was the pattern of results indicative of speed-accuracy trade-offs when compared with the RT data.

To provide a clearer picture of the patterns of performance produced by the three types of competing signal, separate ANOVAs were conducted on the RT data for each competing signal condition.

Forward Speech (Different Ear)

Figure 1 shows the average priming effect produced by biasing contexts presented in the presence or absence of a competing signal in the forward speech condition (different ear). The magnitude of the priming effect was analysed in two-way subject and item ANOVAs with semantic bias and competing signal as within-subjects variables. There was a significant main effect of semantic bias (Subject $F(1,11) = 45.17$, $p < .001$; Item $F(1,29) = 11.28$, $p < .01$), but no significant main effect of competing signal. A significant semantic bias x competing signal interaction emerged for RTs (Subject $F(1,11) = 10.31$, $p < .01$; Item $F(1,29) = 4.54$, $p < .05$). Planned contrasts conducted within a general linear model revealed that inhibition was significantly reduced when the context was presented along with competing forward speech in a different ear, relative to when the context was presented in isolation (Subject $F(1,11) = 20.35$, $p < .001$; Item $F(1,29) = 5.41$, $p < .05$). However, this manipulation did not significantly affect facilitation.

Backward Speech (Different Ear)

Figure 2 shows the average priming effect produced by biasing contexts presented in the presence or absence of a competing signal in the backward speech condition (different ear). The magnitude of the priming effect was analysed in two-way subject and item ANOVAs for RT and accuracy with semantic bias and competing signal as within-subjects variables. There was a significant main effect of semantic bias (Subject $F(1,11) = 48.77$, $p < .001$; Item $F(1,29) = 63.62$, $p < .001$), but no significant main effect of competing signal. The interaction of semantic bias and competing signal did

not approach significance in either the subject or item analysis. Further, planned contrasts revealed no significant effect of this competing signal on facilitation or inhibition.

Forward Speech (Same Ear)

Figure 3 shows the average priming effect produced by biasing contexts presented in the presence or absence of a competing signal in the forward speech condition (same ear). The magnitude of the priming effect was analysed in two-way subject and item ANOVAs with semantic bias and competing signal as within-subjects variables. There was a significant main effect of semantic bias (Subject $F(1,11) = 21.92$, $p < .001$; Item $F(1,29) = 16.02$, $p < .001$), but no significant main effect of competing signal. A significant semantic bias x competing signal interaction emerged (Subject $F(1,11) = 14.24$, $p < .01$; Item $F(1,29) = 4.76$, $p < .05$). Planned contrasts revealed that facilitation was significantly reduced when the context was presented along with competing forward speech in the same ear (Subject $F(1,11) = 22.40$, $p < .001$; Item $F(1,29) = 13.58$, $p<.001$). However, this manipulation did not significantly affect inhibition.

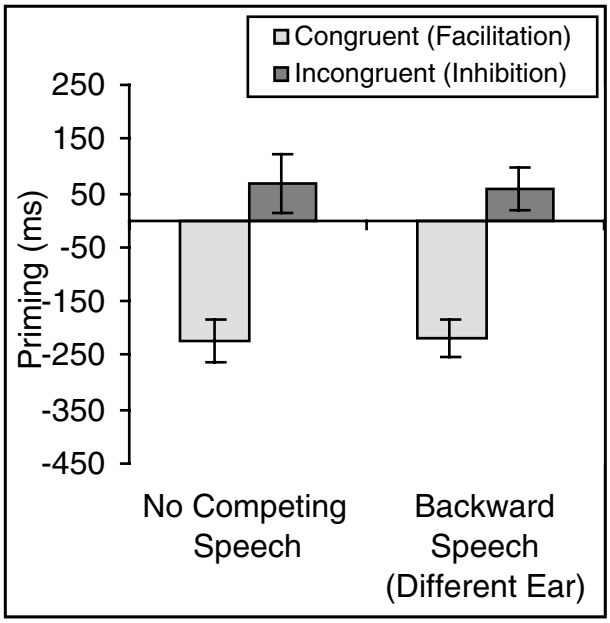


Figure 2. Effect of backward competing speech (different ear). Error bars indicate standard error.

Discussion

These results demonstrate that competing speech modulates the effect of sentence context on the processing of spoken words. Further, the pattern of these effects depends crucially upon the semantic content of the competing signal and the perceptual separability of the competing and target signals. Specifically, competing speech presented in a different ear from the target signal significantly reduced the inhibitory effect of context on incongruent targets, without affecting the facilitatory effect of context on congruent targets. Thus, as predicted, increased attentional demand disrupted only the inhibitory component of the priming effect. However, when backward speech was presented to a different ear, there was no significant effect on facilitatory or inhibitory priming. This finding suggests that it is the semantic content of the competing signal, rather than the signal itself, that increases attentional demand. When forward speech was presented to the same ear, however, the facilitatory effect of context was significantly reduced, whereas the inhibitory effect was unaffected. Thus, when the target signal cannot be isolated from the competing speech, the masking effect of the competing signal disrupts the perceptibility of the target signal, resulting in reduced facilitation.

These results are compatible with previous claims that the facilitatory component of the priming effect reflects the early and automatic processing of lexical-semantic information, whereas the inhibitory component reflects later, more controlled or strategic processes (Faust & Gernsbacher, 1996; Gernsbacher, 1997; Neely, 1991; Utman & Bates, 1998a,b). Future research will examine the implications of these results for language comprehension populations with limited perceptual and attentional capacity, including hearing-impaired and elderly individuals.

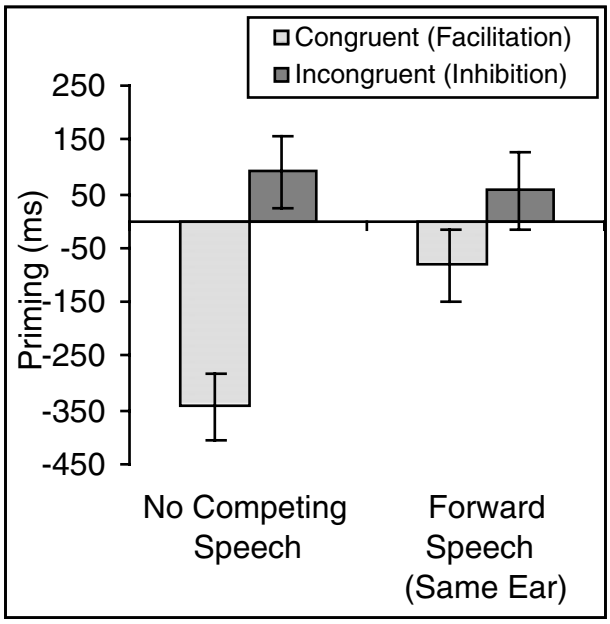


Figure 3. Effect of forward competing speech (same ear). Error bars indicate standard error.

References

Bentin, S., Kutas, M., & Hillyard, S.A. (1995). Semantic processing and memory for attended and unattended words in dichotic listening: behavioural and electrophysiological evidence. *Journal of Experimental Psychology: Human Perception and Performance, 21(1),* 54-67.

Brown, G.D.A. (1984). A frequency count of 190,000 words in ther London-Lund Corpus of English Conversation. *Behavioural Research Methods Instrumentation and Computaters*, *16* (6), 502-532.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, *33A*, 497-505.

Connolly, J.F., Phillips, N.A., Stewart, S.H., & Brake, W.G. (1992). Event-related potential sensitivity to acoustic and semantic properties of terminal words in sentences. *Brain and Language, 43(1),* 1-18.

Downs, D.W., & Crum, M.A. (1978). Processing demands during auditory learning under degraded listening condition. *Journal of Speech and Hearing Research, 21(4),* 702-714.

Duffy, S.A., Henderson, J.M., & Morris, R.K. (1989). Semantic facilitation of lexical access during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15(5),* 791-801.

Faust, M. E., Balota, D. A., Duchek, J. M., Gernsbacher, M. A., & Smith, S. (1997) Inhibitory control during sentence comprehension in individuals with dementia of the Alzheimer type. *Brain and Language*, *57*(2), 225-253.

Faust, M.E., & Gernsbacher, M.A. (1996). Cerebral mechanisms for suppression of inappropriate information during sentence comprehension. *Brain and Language*, *53*(2), 234-259.

Gernsbacher, M.A. (1997). Group differences in suppression skill. *Aging, Neuropsychology, & Cognition*, *4* (3), 175-184.

Garstecki, D.C., & Mulac, A. (1974). Effects of test material and competing message on speech discrimination. *Journal of Auditory Research, 14(3),* 171-177.

Gilhooly, K.J. & Logie, R.H. (1980). Age of acquisition, imagery, concreteness, familiarity and ambiguity measures for 1944 words. *Behaviour Research Methods and Instrumentation*, *12*, 395-427.

Harris, A.E., Benedict, R.H., & Leek, M.R. (1990). Consideration of pigeon-holing and filtering as dysfunctional attention strategies in schizophrenia. *British Journal of Clinical Psychology, 29(1*), 23-35.

Kucera & Francis, W.N. (1967). *Computational Analysis of Present-Day American English*. Providence: Brown University Press.

Marslen-Wilson, W. (1989). Access and intergration: Projecting sound onto meaning. In W. Marslen-Wilson, (Ed.) *Lexical representation and process.* Cambridge, MA: MIT Press.

Marslen-Wilson, W. (1993). Issues of process and representation in lexical access. In G.T.M. Altmann & R. Shillcock (Eds.), *Cognitive models of speech processing: The second Sperlonga Meeting.* Hove, UK: Lawrence Erlbaum.

Meyer, D.E., & Schvaneveldt, R.W. (1976). Meaning, memory structure, and mental processes. *Science, 192(4234),* 27-33.

Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, *11*, 56-60.

Pavio, A., Yuille, J.C. and Madigan, S.A. (1968). Concreteness, imagery and meaningfulness values for 925 words. *Journal of Experimental Psychology Monograph Supplement*, *76* (3, part 2).

Poldrack, R.A., Protopapas, A., Nagarajan, S., Tallal, P., Merzenich, M., Temple, E., & Gabrieli, J.D.E. (1998). *Auditory processing of temporally compressed speech: an fMRI study.* Presented at the Cognitive Neuroscience Society, Fifth Annual Meeting, San Francisco, CA.

Simpson, G.B., Peterson, R.R., Casteel, M.A., & Burgess, C. (1989). Lexical and sentence context effects in word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15(1)*, 88-97.

Slywotzky, A. (1999). *Profit Patterns*. Chichester: John Wiley and Sons.

Stanovich, K.E., & West, R.F. (1983). On priming by a sentence context. *Journal of Experimental Psychology: General, 112(1),* 1-36.

Ulrich, R., & Miller, J. (1994) Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General*, *123*, (1):34-80.

Utman, J. A. & Bates, E. (1998a) Effects of acoustic degradation and semantic context on lexical access: Implications for aphasic deficits. *Brain & Language*, *65* (1), 216-218.

Utman, J. A. & Bates, E. (1998b). *Effects of acoustic distortion and semantic context on lexical access* (Tech. Rep. No. 9803). La Jolla: University of California, San Diego, Center for Research in Language.

Utman, J. A., Dick, F., Prat, C., & Mills, D. (1999) Effects of acoustic distortion and semantic context on event-related potentials to spoken words. Abstract, Cognitive Neuroscience Society, Sixth Annual Meeting, Washington, DC.

Wilcox, R. R. (1992) Comparing the medians of dependent groups. *British Journal of Mathematical & Statistical Psychology*, *45*, (1):151-162.

Wood, N., & Cowan, N. (1995). The cocktail party phenomenon revisited: How frequent are attention shifts to one's name in an irrelevant auditory channel? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21(1),* 255-260.

The consistency of children's responses to logical statements: Coordinating components of formal reasoning

Bradley J. Morris (bmorris@andrew.cmu.edu)
David Klahr (klahr@cmu.edu)
Carnegie Mellon University, Dept of Psychology
5000 Forbes Ave. Pittsburgh, PA 15213

Abstract

Processing of formal statements has three distinct phases: assessing truth-values a priori, requesting evidence, and if requested, evaluating this evidence. Previous investigations of children's ability to process formal assertions have studied each of these phases in isolation, but have not asked whether responses to each processing phase are coordinated into consistent response patterns. This study examined the consistency of 28 third (M= 8.9) and 33 sixth graders (M= 12.0) responses to four types of logical statements: tautologies, contradictions, conjunctions, and disjunctions. The results indicated that third graders' between-item responses are significantly less consistent than sixth graders, that sixth graders are more likely to give correct response on each question phase than third graders, and that development on each question phase may occur independently.

An important milestone in the development of scientific reasoning is the point at which children recognize that some statements are true or false simply because of the their formal structure while others require the gathering and evaluation of empirical evidence. We propose that there are three phases in processing theoretical assertions: 1) evaluating a priori truth values, 2) requesting evidence only if a statement's truth-value is unable to be determined a priori, and 3) correctly mapping evidence to formal states, for statements requiring evidence, so that correct conclusions can be drawn. Thus, a complete account of how children process formal properties requires an examination not only of the development of all three phases in isolation but also of their coordination.

Most previous research in young children's logical reasoning has focused only on one component at a time. That is, research has focused separately on a priori truth-values of statements (Braine & O' Brien, 1998), evidence requests (Osherson & Markman, 1975), or evidence evaluation (Suppes & Feldman, 1971).

This piecemeal focus has led to seemingly contradictory results. For example, preschool children correctly evaluate the a priori truth-values of contradictions as false but incorrectly request evidence for these statements (Osherson & Markman, 1975; Braine & Rumain, 1981). The goal of the present study is to examine children's coordination of all three components involved in processing formal properties.

Logical Statements and Formal Properties

Formal properties describe the relationship between form, evidence and truth-values. A logical statement's formal properties are defined by values on three phases: 1) truth-values before evidence (true, not true, can't tell), 2) whether evidence is necessary (yes, no), and 3) if evidence is necessary, correctly evaluating evidence (true, not true, can't tell). We will focus on the following statement forms: conjunctions, disjunctions, tautologies and contradictions. Each statement type represents different formal properties. The values on each phase distributes the statements into two classes: *logically indeterminate* statements that require evidence to determine their truth-values (hereafter called indeterminate) and *logically determinate* statements that do not require evidence to determine their truth-values (hereafter called determinate) (Suppes, 1957). The statements and components are described in Table1.

Research into young children's processing of formal properties shows mixed evidence for children's competence. While preschool and elementary children show some sensitivity to statements such as contradictions (often evaluating them as false) (Braine & Rumain, 1981) children of the same age tend to request evidence for most problems even when unnecessary and rarely coordinate evaluation and evidence requests correctly (Osherson & Markman, 1975; Morris & Sloutsky, 1999; Ruffman, 1999).

To address this limitation, we examined children's response patterns to determine the degree of between-phase dependence as an index of processing competence. That is, if children correctly process formal properties, then we would expect correct responses in each of the necessary phases. For example, correctly responding to a contradiction requires processing two phases: an a priori evaluation of "not true" and denying a request for evidence. However, a child who lacks such an understanding may not only err on individual question phases (e.g., "true" a priori evaluation) but may fail to link response phases (e.g.,

Table 1- Comparison of Logical Statements by Dimension

	Statement	Example	A Priori?	Evidence Necessary	Evidence/Form Mapping
Determinate	Tautology	The shape is a circle or the shape is not a circle	True	No	All True states
	Contradiction	The shape is a triangle and the shape is not a triangle	False	No	No True States
Indeterminate	Conjunction	The shape is a square and the shape is not red	Can't Tell	Yes	1 True State (square, not red)
	Disjunction	The shape is a square or the shape is not blue	Can't Tell	Yes	2 True States (square) (not blue)

denying request for evidence). In examining this question, we implicitly examined a related question: Are children's responses from a single dimension of formal properties (e.g., a priori evaluations) diagnostic of their overall processing competence?

To examine these research questions, we presented a 'game' to a group of third and sixth grade children in which they were asked to evaluate 16 logical statements. For each statement, each child was asked up to three questions ("phases") per statement: a priori evaluation, a request for evidence, and, if evidence was requested, evidence evaluation. Each child's response to each question phase was coded as correct or incorrect then these question phases were compiled into a response pattern for each of the 16 statements. Response patterns were aggregated and compared to levels predicted by 'chance' decisions at each question phase.

Method

Participants

A total of 28 third graders (8.5-9.5 years, *M*= 8.9) and 33 sixth graders (11.3-13.6 years, *M*= 12.0) participated. The children were recruited from three public schools in Pittsburgh, PA.

Materials

There were four examples of each of four statement types: tautology, contradiction, disjunction, and conjunction. 'Evidence' was provided in the form of a picture (approximately 5"x7") that displayed information related to each statement card. For example, given the statement "The circle is red AND the circle is not red" the picture displays a red circle. Four additional cards and pictures were used for the warm-up tasks. The statements and evidence were designed to allow us to differentiate among response patterns.

Procedure

The experiment was conducted in a single 15-20 minute session that included two segments: a warm-up and an experimental segment. Each participant was tested individually. In the warm-up segment, the interviewer read a set of instructions explaining the game's purpose. The warm-up period lasted approximately five minutes and consisted of four trials. All instructions and statements for the experimental segment were read to each participant and repeated if requested. Two cards were placed in front of each child: a statement card (face up) and a picture card (face down). The order of presentation was counterbalanced across participants. Each child was given 16 trials, with 1 statement per trial.

For each trial, the following procedure was used. The statement card and picture were placed in front of the child. The child was read the statement card. (Q1) The child was then asked the first question phase: An evaluation of truth status of the statement before evidence (*a priori* evaluation): "Is this statement True, Not True, or Can't Tell. (Q2) After the child responded to the statement, they were asked a second question: Do you need to see the picture to help figure out the sentence? Yes or No. (Q3) If the child requested to see the picture, then they were shown the picture and asked the following: "Now that you have seen the picture, is the statement True, Not True, or Can't Tell."

Coding

Compiled Phase Analysis (CPA) Traditionally, children's responses to each phase of questioning have been analyzed independent of their responses to the other phases. We suggest an analysis of children's response patterns, that is, the frequency of the different types of sequential responses that children generated as they moved through each of the three processing phases for each statement. Because each pattern represents the compilation of

responses across the three phases, we call this the Compiled Phase Analysis (CPA).

The set of all possible response patterns – both correct and incorrect – is listed in Table 2. For example, as shown in the top row of Table 2, if a child's responses to a Contradiction were A priori- "False" and Evidence Request- "No", we would code that pattern as C-C because each phase was answered correctly. But many other patterns could – and did - occur. For example, one erroneous response pattern for a contradiction is: A priori- "False", Evidence Request- "Yes" and, Evidence Evaluation- "True," which would be coded as C-I-I, (Table 2 last row of upper section) because the answer to the first phase is correct, but the other two are incorrect. This pattern demonstrates correct a priori evaluation, but incorrect evidence request and evaluation. Another erroneous response pattern for a contradiction is A priori- "Can't Tell", Evidence Request- "Yes", Evidence Evaluation- "False". (Coded as I-I-C). This pattern reveals a different type of misunderstanding: one in which, because the contradiction is not recognized as such, evidence is (incorrectly) requested, but then correctly evaluated. The other patterns listed in Table 2 imply other types of errors. For the remainder of this section, we will look more closely at the distribution of these patterns, without further discussing their underlying implications.

These distributions are informative because a child responding at random to each phase could have generated any particular response pattern. By analyzing the relative frequency of different response patterns – in particular the extent to which they deviate from what would be expected from a random responder - we can begin to understand whether children are responding consistently, albeit erroneously, to these statements or whether their responses to each phase are random, and based only on guessing.

Examining Response Patterns In this section we investigate the extent to which children's response patterns deviate from a randomly generated set. Chance values are based on the assumption that, for each phase, each alternative response is equally likely and is independent of all other phases. Response patterns consisted of either three questions (if evidence was requested) or two questions (if evidence was not requested). The first, a priori evaluation, has three possible responses: "true," "not true" or "can't tell," only one of which is correct, with probability 1/3. The second phase, request for evidence, had two possible responses: "yes or no". Thus the probability of correctly answering this question was 1/2. The third phase, evidence evaluation, occurred only if evidence had been requested (and obtained). Like the first question, this question had three possible responses, with the probability of the one correct response being 1/3.

These probabilities were used to generate the expected chance distribution of the different response patterns for determinate and indeterminate statements. The expected and observed results for each type of pattern are displayed in Table 2 as counts. Also displayed in Table 2 are the observed proportions of each type of pattern for each grade level.

For determinate and indeterminate statements, there were 6 possible response patterns, one correct and five incorrect. Chance values were calculated by computing the probability of choosing a correct or incorrect response on each question phase. For example, for the determinate pattern (I-I-I), the probability of an incorrect a priori response is 2/3, an incorrect evidence request is 1/2, and an incorrect evidence evaluation is 2/3. The conditional probability of selecting this pattern of responses is 2/3*1/2*2/3=2/9. Once the conditional probability was calculated for each possible response pattern, we calculated the total number of patterns we would expect by chance given the number of responses in the data set (229, 3rd graders). So the numbers in Table 2 reflect the number of times we would expect to see each response pattern if a child simply 'flipped a coin' at each decision point.

Finally, the response pattern analysis also indicates the degree to which children correctly process individual response phases. That is, if children err in processing one question phase (e.g., a priori evaluations only) then errors should focus on those patterns that indicate correct response on two phases and errors on one phase. Further, the overall distribution of responses should deviate from chance on patterns that reflect errors on this particular response phase. Such evidence would support the notion of independent developmental trajectories for each question phase.

Results

Configurational Frequency Analysis

We conducted a configurational frequency analysis to compare the observed and expected values from the CPA analysis (von Eye, 1990). This analysis uses assumptions similar to a chi-square analysis to compare the distribution of expected and observed response frequencies. This analysis controls the overall significance level by using the Bonferroni adjustment. The results provide a

significance level for the difference between the observed and expected values. When these differences are significant the CFA indicates two classifications: types (in which the observed value is significantly higher then the predicted value) and antitypes (in which the observed value is significantly lower then the predicted value). The analysis was conducted using the 'CFA program for 32 bit operation systems' (von Eye, 1998). The results of the analysis determinate the significance levels indicated in Table 2. Recall that for each of the possible CPA patterns, there is one correct and five incorrect patterns for both determinate and indeterminate statements.

For indeterminate statements, sixth grader's correct CPA responses deviate significantly from chance, while third graders CPA responses do not differ significantly from chance. Sixth graders are below chance on one incorrect pattern. Conversely, third grader's response patterns were above chance on one pattern- I-C-C- and below chance on two patterns- C-I and I-I. This suggests that children requested evidence for most problems failing to distinguish when evidence was unnecessary and often failed to assign correct truth-values before evidence. For determinate statements, neither third nor sixth graders' correct patterns were above chance. Of note however, is that both sixth and third grader's response patterns were significantly above chance for one incorrect pattern- I-C- suggesting that they were not processing a priori and evidence request responses dependently. Third graders were above chance for one incorrect pattern- I-I-I indicating that they erred on all question phases.

Finally, an examination of the types (observed levels are significantly above those predicted by chance) and anti-types (observed levels are significantly below those predicted by chance) supports the notion of independent developmental trajectories for each question phase. Third graders response patterns demonstrated a lack of understanding of the necessity of evidence often erroneously requesting evidence for determinate statements. Few third (or sixth) graders failed to request evidence for indeterminate statements (I-I and C-I patterns) suggesting that in general, children erred on the side of evidence requests.

For determinate statements, third and sixth graders also made fewer I-C patterns than expected by chance. Thus, children rarely made an error on a priori evaluations and were correct on evidence requests suggesting that the former is better established in third graders than the latter. Overall, children's patterns suggested that errors on evidence evaluation were the least frequent while errors on evidence requests were the most frequent.

Correct Response Patterns

In order to compare the number of correct responses by age and type a 2 (grade) X 2 (statement type: determinate v. indeterminate) ANOVA was performed with grade as a between-subjects factor and statement type as a within-subjects factor. Each correct pattern was scored as 1 while each incorrect pattern was scored as 0. Results indicated that sixth graders gave significantly more correct responses on both determinate F (1, 114) = 39, $p < .001$ and determinate statements F (1, 114) = 42, $p < .001$ than third graders.

CPA: Summary

The Compiled Phase Analysis examined the degree to which children are correctly processing each question phase. The formal properties of each statement type require a particular response pattern in which each question phase is dependent on the previous response. The CPA compared the number of responses due to chance responding on each phase to the number of observed responses. The results suggest 1) a high amount of variability in all children's between phase processing, 2) sixth graders were more likely than third graders to correctly process all phases, and 3) each processing phase develops independently. However, while the CPA indicated phase to phase dependencies, it does not reveal the specific strategies that could produce these patterns (see Morris & Klahr, in review for such as analysis).

Table 2- Predicted and Observed Compiled Response Patterns

Type of response patterns[a]					*Number of responses of each type*[b]			
Statement Type	*A Priori*	*Evidence Request*	*Evidence Evaluation*	*Pattern Code*	*Third Grade*		*Sixth Grade*	
Determinate	**Correct**	**Correct**	**N/A**	**C-C**	15	(38)	85	(44)
	Incorrect	Correct	N/A	I-C	12	(74)+	21	(87)+
	Incorrect	Incorrect	Incorrect	I-I-I	96	(50)*	86	(58)
	Incorrect	Incorrect	Correct	I-I-C	51	(25)	40	(29)
	Correct	Incorrect	Correct	C-I-C	27	(14)	17	(16)
	Correct	Incorrect	Incorrect	C-I-I	23	(25)	14	(29)
					224		263	
Indeterminate	**Correct**	**Correct**	**Correct**	**C-C-C**	35	(13)	110	(16)*
	Correct	Correct	Incorrect	C-C-I	25	(25)	40	(29)
	Correct	Incorrect	N/A	C-I	1	(38)+	3	(44)+
	Incorrect	Correct	Incorrect	I-C-I	57	(48)	22	(58)
	Incorrect	Correct	Correct	I-C-C	78	(25)*	40	(29)
	Incorrect	Incorrect	N/A	I-I	27	(74)+	47	(86)
					223		262	

a. Responses are coded as being 'correct' or 'incorrect' compared to optimal response patterns. Correct response patterns are in bold.

b. Entries indicate the total number of individual response patterns summed over all children's responses in each grade. A total of 28 3rd graders and 33 6th graders were given 8 determinate and 8 indeterminate statements. Numbers in parentheses indicate the number of responses expected from a random response pattern.

* Indicates a type (significantly above chance at the $p < .05$ level) while + indicates and anti-type (significantly below chance at the $p < .05$ level).

Discussion

We have argued that one previously overlooked component of logical reasoning is the coordination of individual responses into consistent response patterns that reflect an understanding of formal properties. The present study attempted to examine simultaneously children's responses to each component of formal properties (operationalized as component question phases) and to compare the distribution of children's responses to what would be expected if children answered each question phase as if they were independent of each other.

The results indicated that sixth graders' compiled response patterns demonstrated a greater degree of question phase dependence than the response patterns of third graders. When the aggregated patterns were compared to chance, sixth graders' responses were significantly above chance for 'correct' patterns on indeterminate statements while third graders patterns were above chance only for a specific error pattern for indeterminate problems (I-C-C). Overall, third graders response patterns deviated from levels predicted by 'chance' responding only on incorrect response patterns.

The results suggest that each of the component properties of formal properties has an individual developmental course and that once each component property is established, the individual components must then be coordinated. The distributions of response patterns suggest the following: 1) children are likely to err when requesting evidence demonstrated by high levels of incorrect evidence

requests for determinate statements and low levels of evidence refusal for indeterminate statements, 2) sixth graders often evaluated evidence correctly, even when such evidence is unnecessary. Taken together, the evidence suggests that correct processing on all phases is not present early in development and that correct processing may occur at different times for each phase.

While sixth grader's performance was significantly better than third graders, they still erred on a large number of statements suggesting that they do not correctly process formal properties.

Finally, the large amount of variance in children's response patterns suggests that a focus on one component of formal properties is not diagnostic of children's overall understanding of formal properties. That is, children's responses on one component may erroneously suggest competence (or lack of competence) when only partial competence exists.

Acknowledgements

This work was supported in part by a postdoctoral fellowship to the first author from a NICHD (HD08550) and in part by a grant to the second author from NICHD (HD25211). Thanks to Anne Siegel for data collection, coding and analysis and Jen Schnakenberg and Amy Masnick for helpful comments on a previous draft.

References

Braine, M., & Rumain, B. (1981). evelopment of comprehension of "Or": Evidence for a sequence of competencies. *Journal of Experimental Child Psychology, 31,* 46-70.

Braine, M. & O'Brien, D. (Eds.). (1998). *Mental Logic*. NJ: Lawrence Erlbaum.

Morris, B. J., & Sloutsky, V. (1999). Developmental differences in young children's solutions of logical and empirical problems. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the twenty-first annual conference of the Cognitive Science Society* (pp. 432-437). Mahwah, New Jersey: Lawrence Erlbaum.

Morris, B.J., & Klahr, D. (in review). The role of evidence in children's reasoning strategies.

Osherson, D., & Markman, E. (1975). Language and the ability to evaluate contradictions and tautologies. *Cognition, 3*(3), 213-226.

Ruffman, T. (1999). Children's understanding of logical inconsistency. *Child Development, 70 (4)*, 872-886.

Suppes, P. (1957). *Introduction to Logic.* New York: Van Nostrand.

Suppes, P. & Feldman, S. (1971). Young children's comprehension of logical connectives. *Journal of Experimental Child Psychology, 12,* 304-317.

von Eye, A. (1990). *Introduction to configurational frequency analysis: The search for types and antitypes in cross-classifications.* Cambridge, UK: Cambridge University Press.

von Eye, A. (1998). CFA program for 32 bit operation systems. *Methods of Psychological Research - online, 3,* 1 - 3.

Working-memory modularity in analogical reasoning

Robert G. Morrison (morrison@psych.ucla.edu)
Keith J. Holyoak (holyoak@lifesci.ucla.edu)
Bao Truong (bt@ucla.edu)

University of California, Los Angeles
Department of Psychology, BOX 951563
Los Angeles, CA 90095-1563 USA

Abstract

We present several experiments using dual-task (DT) methodology to explore use of working memory (WM) during analogical reasoning. Participants solved verbal and figural analogy problems alone or while performing articulatory suppression (AS), spatial tapping (ST) or verbal random generation (VRG). As in other studies of relational reasoning we found that VRG disrupted both verbal and figural analogy performance. In addition, we found disruption of analogy performance by WM slave system distractors (i.e., AS and ST) consistent with the dominant modality of the analogy task. These findings are discussed with respect to Baddeley's model of WM and other studies of WM involvement in relational reasoning.

Introduction

Central to the ability to reason by analogy is the ability to form and manipulate mental representations of relations between objects and events. For instance, in a verbal analogy such as:

BLACK:WHITE::NOISY:QUIET

the reasoner needs to form mental representations of the relation between BLACK and WHITE (black is the opposite of white) and map it to the second pair in order to verify that the analogy is appropriate. Thus, BLACK:WHITE is mapped to NOISY:QUIET and the analogy is successfully solved. It has long been assumed that this type of process requires the use of WM (cf., Baddeley & Hitch, 1974); however, until recently relatively little attention has been given to how WM limits affect analogical reasoning (Halford et al., 1994; Hummel & Holyoak, 1992; Hummel & Holyoak, 1997; Keane, Ledgeway, & Duff, 1994). In the present paper we report experiments using dual-task (DT) methodology (employed extensively by Baddeley, 1986) to study the involvement of the various modules of WM in analogical reasoning.

Baddeley's (Baddeley, 1986; Baddeley & Hitch, 1974; Baddeley & Logie, 1999) model of WM has dominated cognitive accounts of short-term memory for nearly three decades. The model consists of three components: the Phonological Loop (PL), the Visuo-Spatial Sketchpad (VSSP), and the Central Executive (CE). In Baddeley's model the PL and VSSP are modality-specific slave systems that are responsible for maintaining information over short periods of time. Baddeley (Baddeley & Hitch, 1974) originally conceived of the CE to account for functions of WM not performed by the PL and VSSP; however, Baddeley (1986) later embraced Norman and Shallice's (Norman & Shallice, 1986) Supervisory Attentional System as a possible model of the CE. Most recently, Baddeley (1996) has segmented the CE in an attempt to study its component processes. From this perspective the CE is responsible for (1) the capacity to coordinate performance on 2 separate tasks, (2) the capacity to switch retrieval strategies as reflected in random generation, (3) the capacity to attend selectively to 1 stimulus and inhibit the disrupting effect of others, and (4) the capacity to hold and manipulate information in long-term memory, as reflected in measures of WM span (Baddeley, 1996 p. 5). Baddeley suggests that the CE manages the work of WM while the slave systems actually maintain the information.

Also central to Baddeley's model is the concept of limited capacity. The slave systems and the CE share this limited capacity, such that increasing CE functioning would reduce the capacity of either the PL or VSSP to maintain information; however, there is evidence that each system may have its own limits as well (e.g., the PL capacity is limited by the amount of information that can be subvocally cycled in approximately two seconds).

Evidence for a multi-module WM system is copious, coming from both the cognitive and neuropsychology literatures. However, relatively little attention has been paid to the implications of WM for relational reasoning particularly analogical reasoning. Review of the functions of the CE as outlined above suggests that the CE should be critical for relational reasoning. Experimental evidence has confirmed this hypothesis for deductive reasoning, with random generation (e.g., Baddeley & Hitch, 1974; Gilhooly, Logie, Wetherick, & Wynn, 1993; Klauer, Stegmaier, & Meiser, 1997) and memory load (e.g., Baddeley & Hitch, 1974; Gilhooly et al., 1993; Toms, Morris, & Ward, 1993) both interfering

with performance. Klauer et al. (1997) found that random generation interfered with spatial reasoning (transitive inference), and Waltz, Lau, Grewal, and Holyoak (2000) found that performing VRG or maintaining a concurrent memory load discouraged participants from using relational mappings in a task that can be solved via either featural or relational similarity (see Markman & Gentner, 1993, for a task description). What is not clear from these studies is what aspects of the DTs actually cause the interference in relational reasoning. At the very least, random generation involves task switching, memory insertion and storage, and relational binding of numbers with temporal location; whereas maintaining a concurrent memory load involves memory insertion and storage. Both tasks are very demanding on WM resources.

It is also not clear to what extent the WM slave systems are important for reasoning, particularly in situations where all the information is available visually to the reasoner. Gilhooly et al. (1993) and Toms et al. (1993) found no effect of PL- or VSSP-based DTs on propositional reasoning, while Klauer et al. (1997) found a small effect of articulatory suppression (AS; a PL secondary task) on reasoning latencies. It is important to note that in each of these propositional reasoning tasks all information necessary to complete the task was perceptually available in the task. For example, a propositional reasoning problem such as:

There is either a circle or a triangle.
Therefore, there is no triangle.

requires only the information presented to answer the problem. In contrast, a transitive inference problem such as:

The circle is to the right of the triangle.
The square is to the left of the triangle.
Therefore, the square is to the left of the circle.

requires the reasoner to generate a new proposition based on the information presented (i.e., left-of (square, circle)).

Similarly, Waltz et al. (2000) found that performing AS while performing the Markman and Gentner similarity task discouraged participants from using relational correspondences just as VRG did. A recent replication of this result in our lab showed that ST had an effect similar in magnitude to AS. Like the transitive inference task described previously, in order to make a relational choice propositions not immediately obvious from the stimuli must be generated. This characteristic is a hallmark of analogical reasoning. Thus, it is not clear at present to what extent the slave systems of WM are necessary for relational reasoning. It is likely that the modality and quantity of information that must be retrieved and relationally bound in order to perform a reasoning problem will determine which WM slave systems will be necessary.

Methods

To explore to what extent the various modules of WM are recruited in analogical reasoning, participants performed several relational reasoning tasks while performing one of several DTs. Participants in the AS condition were instructed to say the English non-word zorn once each second. Another group in the ST condition was instructed to tap four red dots in a clock-wise pattern one dot each second. Participants in the VRG condition were instructed to say a random digit from 0 to 9 once each second. A fourth group of participants served as controls, performing only the primary reasoning tasks. 96 undergraduate students from the University of California, Los Angeles participated in the study in exchange for course credit.

Verbal Analogy

In the verbal analogy (VA) task participants verified A:B::C:D analogies such as: BLACK:WHITE:: NOISY:QUIET (i.e., participants answered TRUE or FALSE). Analogy problems were based on those developed by Sternberg and Nigro (1980). A:B pairs were related by one of five common relations (antonyms, synonyms, category members, functions, or linear ordering). In TRUE problems, C:D pairs shared the same relation as the A:B pairs but were from a different domain than the A:B pair (e.g., color vs. sound). We created FALSE problems by substituting a D term that was related to C in a different way (e.g., linear-ordered (noisy, noisier) instead of opposite-of (noisy, quiet)).

Figural Analogy

In the figural analogy task (FA) participants verified A:B::C:D analogies based on Sternberg's (1977) People Piece Analogy (PPA) task. In PPA each item was a cartoon character that possessed one each of four binary traits (male/female, black/white, tall/short and fat/thin). TRUE analogies showed the same changes in traits between the A:B pair and the C:D pair as well as between the A:C pair and the B:D pair. Problems of varying degrees of relational complexity (RC, cf. (Halford, Wilson, & Phillips, 1998) were constructed based on the number of traits that were manipulated. RC=1 problems had only one trait manipulated across either the A:B or A:C pair. Thus, RC=1 problems were semi-degenerate, with only two repeated characters making up the entire analogy (see Figure 1a). RC=2 and RC=4 problems had either 1 or 2 traits manipulated across both the A:B and A:C pairs (for a total of either 2 or 4 total relations). Thus, RC=2 and RC=4 problems were non-degenerate, consisting of four unique characters in each problem (see Figure 1b). We created FALSE items by changing the Identity of one trait in the

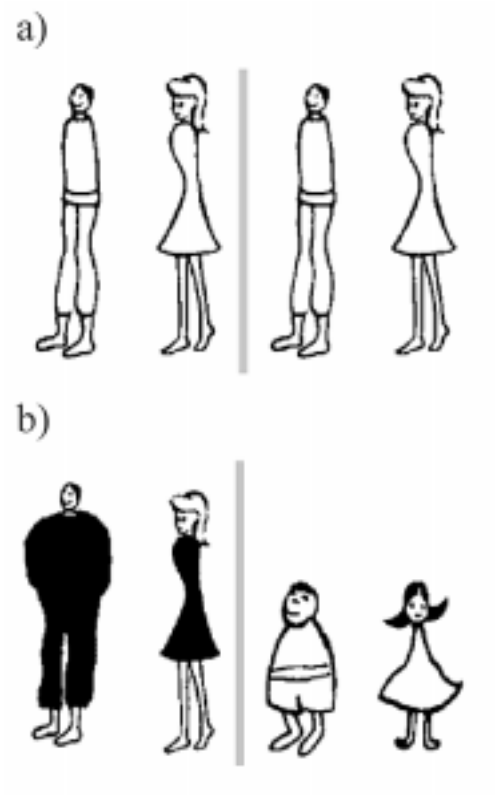


Figure 1: PPA figural analogy problems. a) semi-degenerate, RC=1, b) non-degenerate, RC=4.

fourth character so that it was not analogous.

Procedure

Reasoning problems were presented on a computer screen and participants indicated their response by pressing either a left or right foot pedal. Prior to beginning an experimental block, participants practiced their DT alone, the reasoning problems alone, and then practiced the two tasks together. Participants in the control group practiced the same total number of reasoning problems as participants in the DT groups. Reasoning problems were presented in three one-minute blocks. PPA problems were presented in blocks of increasing RC. The computer recorded analogy RT and accuracy as well as the frequency at which participants performed their DT.

Each block began with the participant pressing the right foot pedal. The participant was instructed to begin their DT. After 5s the first analogy problem appeared on the monitor. When the first problem appeared the experimenter began to hit a key each time the participant performed their DT. In this way the actual frequency and spacing of DT performance was recorded. After 60s a prerecorded voice told the participant to stop both tasks. The next block began after a 30s delay. After the final block the participant received instructions on the next task in the testing battery. The order of tasks was counterbalanced across participants.

Results

We predicted, as in past studies of both deductive (Gilhooly et al., 1993; Klauer et al., 1997) and analogical (Waltz et al., 2000) reasoning, that VRG would interfere with reasoning in both analogy tasks. We also predicted, as in a past study of transitive inference (Klauer et al., 1997) and analogy (Waltz et al., 2000), that DTs that interfered with WM slave systems corresponding to the modality of the task would interfere with performance. Thus, we expected that AS would interfere with VA performance and that ST and possibly AS (because of a verbal strategy frequently employed during PPA solving) would interfere with FA performance.

We analyzed both reasoning and DT performance from both the VA and FA tasks with between-subjects analysis of variance (ANOVA). In addition, we examined reasoning task performance by comparing control group performance to each of the DT groups using single DF planned comparisons.

Verbal Analogy

VA task performance is summarized in Figure 2. A between-subjects ANOVA revealed a reliable effect of DT type on accuracy (d-prime); F (3,92) = 4.5, MSE = .44, p = .005. Planned comparisons showed that AS and VRG had reliable effects on VA accuracy; t (46) = 3.7, p = .003 and t (46) = 2.8, p = .008, respectively. ST did not have a reliable effect on VA performance, t (46) = 1.1, ns. We conducted a similar analysis on RTs for the VA results. An ANOVA revealed a nearly reliable effect of DT type on VA RT; F (3,92) = 2.3, MSE = 979432, p = .085. Planned comparisons showed that VRG had a reliable effect on VA RT; t (46) = 2.3, p = .025. AS and ST did not have a reliable effect on VA RT; t (46) = .35, and t (46) = .96, respectively, both ns. DT data were analyzed using two metrics. First, a measure of DT frequency (mDT) was calculated for each subject (mean time between repetitions in ms). Second, a standardized measure of DT variance (vDT) was calculated for each subject (SD of time between repetitions divided by mDT). Participants performed AS (M =789 ms) and VS (M = 612 ms) faster than VRG (M = 1165 ms); t (69) = 4.4, $p < .001$. A second planned comparison showed that participants performing VRG were more variable in their performance than those performing AS or ST, even when the variance was corrected for the difference in mDT (vDT); t (69) = 3.2, p = .01. Thus, results for the VA task suggest that both the phonological loop (AS DT) and central executive (VRG DT) are important for performance of verbal analogies, with VRG producing a greater effect.

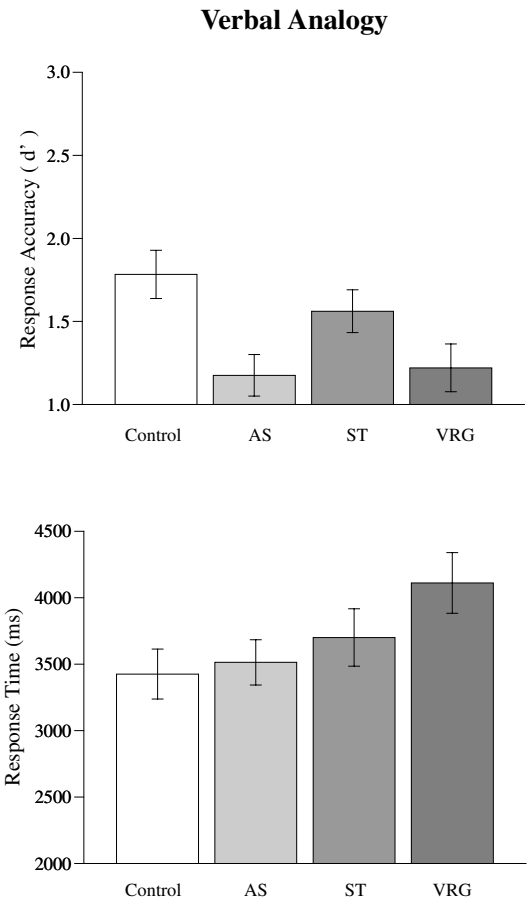


Figure 2: Accuracy and RT performance for verbal analogy under different dual-task conditions. Error bars reflect SEM.

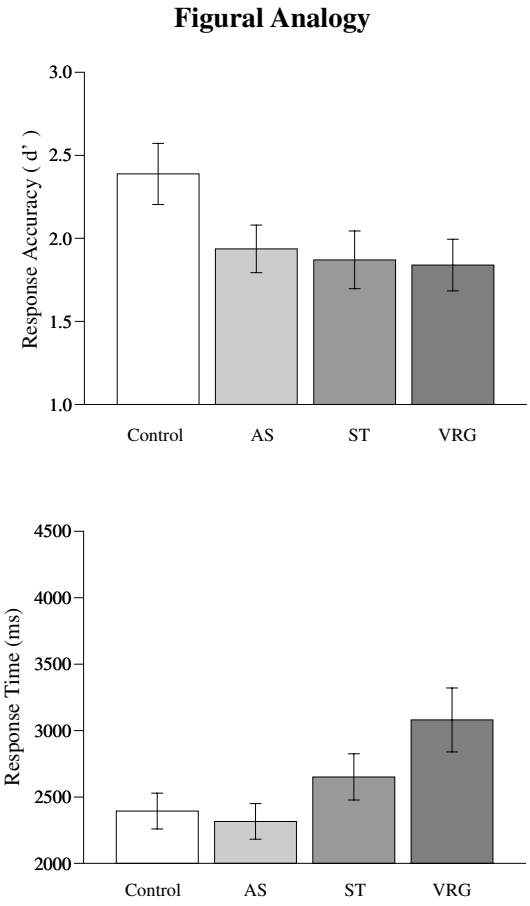


Figure 3: Accuracy and RT performance for People Pieces analogy under different dual-task conditions. Error bars reflect SEM.

Figural Analogy

FA task performance is summarized in Figure 3. A between-subjects ANOVA revealed a reliable effect of DT type on accuracy (d-prime); F (3,92) = 3.1, MSE = .65, p = .032. Planned comparisons showed that ST and VRG had reliable effects on FA accuracy; t (46) = 2.2, p = .036 and t (46) = 2.7, p = .01, respectively. AS had a marginal affect on FA accuracy; t (46) = 1.9, p = .059. We conducted a similar analysis on RTs for the FA results. An ANOVA revealed a reliable effect of DT type; F (3,92) = 3.8, MSE = 748368, p = .013. Planned comparisons showed that VRG had a reliable effect on RT for the FA task; t (46) = 2.5, p = .017. AS and ST did not have reliable effects; t (46) = 1.2 and t (46) = .41, respectively, both ns. Participants performed AS (M =1030 ms) and VS (M = 849 ms) faster than VRG (M = 1626 ms); t (69) = 2.6, p = .01. A second planned comparison showed that participants performing VRG were more variable in their performance than those performing AS or ST, even when the variance was corrected for the difference in mDT (vDT); t (69) = 2.5, p = .02. Thus, results for the FA task suggest that both the visuospatial sketchpad and central executive are important for performance of figural analogies, with the phonological loop perhaps playing a more minor role.

Discussion

In this study we have shown that WM slave systems can be recruited in the service of analogical reasoning and that the specific WM slave systems involved depend of the dominant modality of the task. This result agrees with a previous study of the affect of DTs on analogical reasoning (Waltz et al., 2000) and also of a similar study involving transitive inference (Klauer et al., 1997). All of these tasks have in common the need to generate propositional structures based on the information present in the problem.

In addition to the involvement of the WM slave systems we found robust effects of VRG on both verbal and figural analogies. This result is consistent with a growing body of findings for both deductive and inductive reasoning.

Multimodal vs. Unimodal Working Memory

Discussion of WM has traditionally been divided into two camps camps that are frequently divided by the Atlantic Ocean. The multi-modal camp (centered to the east of the Atlantic) has typically relied on DT methodologies, and results from neuropsychology and more recently neuroimaging. The uni-modal camp (centered to the west of the Atlantic) favors WM-span measures used as probes to investigate individual differences in language and reasoning. The current results, while not inconsistent with the capacity limits that are central to the uni-modal models, require a multi-modal model for a complete interpretation.

On first consideration, DTs such as AS or ST could simply require less WM resources than tasks such as VRG. Inductive reasoning tasks that require retrieval of semantic information and/or generation of additional propositions in WM may simply be more load intensive than propositional reasoning tasks in which all of the information necessary to solve the problem is perceptually available. Thus, AS and ST interfere with inductive reasoning and not propositional reasoning (at least the simple propositional reasoning problems typically used in DT studies). This account predicts that slave system tasks should interfere with reasoning less than VRG and also predicts no dissociation of PL or VSSP DTs if AS and ST interference is simply load dependent then the modality of the resource drain should not matter. The results of Waltz et al. (2000) argue against the weak form of this interpretation, in that AS and VRG produced equal interference in the analogy task. However, it is possible that the analogy measure used in their task (which shows a robust individual difference most likely not related to WM capacity) may not have been sensitive enough to pick up the differences in resource demand caused by AS and VRG. Also, performance on the secondary tasks was not assessed in that prior study.

The results reported here--showing a dissociation in slave system DT interference across analogy tasks of different modality--rule out this interpretation and argue for a multi-modal WM system that requires separate phonological and visuospatial systems. Specifically, the finding of strong interference by AS in the VA task with no corresponding ST interference argues that the PL is necessary for verbal analogy, while the VSSP is not. In contrast, the stronger interference of ST in the predominantly visual FA task compared to PL shows the opposite pattern of interference.

It is not clear from these results, however, what role the slave systems play in analogy. One possibility is that they are used to maintain relational information while it is organized into the propositional structures necessary for further relational processing. In this view, AS and ST DTs interfere with activation of the semantic or visual information necessary to solve the analogy task. This interpretation is consistent with Baddeley's view of the slave systems if one considers the role of the PL and VSSP to be maintenance of representations via continual firing of their mental representations in long-term memory (LTM), a conception proposed by Fuster (1997).

The Role of the Central Executive in Reasoning

One criticism of the multi-modal WM model has been the amorphous nature of the CE. However, a general consensus among researchers is beginning to emerge: the CE is viewed as important for task switching, inhibition of internal representations or prepotent responses, and the activation of information in LTM during an activity that requires the active manipulation of material. All of these functions appear to be critical for higher-level cognition--particularly relational reasoning. What this consensus fails to provide is a detailed account of how the CE actually performs relational reasoning.

Hummel and Holyoak (1997) proposed a model of how the CE may perform relational reasoning. This model, LISA (Learning and Inference with Schemas and Analogies), is an artificial neural-network model of relational reasoning. LISA uses synchrony of firing to bind distributed representations of relational roles (e.g., the roles of *opposite-of* (X, Y)) to distributed representations of their fillers (e.g., *black* and *white*). The process of "thinking about" a proposition, such as *opposite-of* (black, white), entails keeping separate role-filler bindings (e.g., those for black and those for white) firing *out* of synchrony with one another. According to LISA, WM is therefore necessarily capacity-limited: It is only possible to keep a finite number of role-filler bindings simultaneously active and out of synchrony with one another. The synchronized (and de-synchronized) patterns of activation representing propositions in LISA serve as the basis for memory retrieval, analogical mapping, analogical inference and schema induction. That is, all the operations of WM depend critically on the role-filler bindings in WM. As such, an important component of the "job" of the CE is to control which patterns enjoy the "privilege" of remaining active and mutually desynchronized. This process requires no homunculus to operate; rather, it is governed simply by the way that relational information is structured in LTM and the extent to which different mental representations are relationally similar.

According to LISA, a second function of the CE is to keep track of the correspondences between elements of the source and elements of the target (see Hummel & Holyoak, 1997). Algorithmically, LISA accomplishes this function by monitoring which units in the source fire in synchrony with which in the target. Hummel and Holyoak assume that this "keeping track" is performed by neurons in prefrontal cortex with rapidly-modifiable

synapses (e.g.,Asaad, Rainer, & Miller, 1998; Fuster, 1997), and thus needs no greater executive control.

If these are the roles of the CE in relational reasoning, then why does VRG so potently interfere with reasoning? We argue that VRG requires exactly the same operations as relational reasoning. To produce a random stream of numbers it is important not only to know what numbers one has recently said (e.g., 3,8,2), but also the order in which one said them (e.g., 3,8,2,8,2,3 seems more "random" than 3,8,2,3,8,2; Baddeley, 1966, noted that as VRG performance breaks down di- and tri-grams start to emerge in the number stream). That is, it is necessary to bind the numbers to their serial position. According to LISA, VRG consumes exactly the kind of binding resources as the binding and mapping of relational information in WM. As a result, VRG disrupts analogical reasoning and other forms of relational reasoning.

Acknowledgments

This project was supported by NSF grant SBR-9729023 and NIH training grant MH-1992605. We thank Michael Gardner for providing the verbal analogy problems, Alex Antoniu, Arden Ash, Steven Pranoto, Scott Rosnick, Maria Selah, Ji Son, Roshaun Stanisai and Leah Swalley for excellent technical assistance and John Hummel, Dan Krawczyk and others in the Symbolic Neural Computation group at UCLA (www.sync.psych.ucla.edu) for helpful discussions.

References

Asaad, W. F., Rainer, G., & Miller, E. K. (1998). Neural activity in the primate prefrontal cortex during associative learning. *Neuron, 21*, 1399-1407.

Baddeley, A. (1996). Exploring the central executive. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 49*(1), 5-28.

Baddeley, A. D. (1966). The capacity for generating information by randomization. *Quarterly Journal of Experimental Psychology, 18*, 119-129.

Baddeley, A. D. (1986). *Working memory*. Oxford, UK: Oxford University Press.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8). New York: Academic Press.

Baddeley, A. D., & Logie, R. H. (1999). Working memory: The multiple component model. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 28-61). New York: Cambridge University Press.

Fuster, J. M. (1997). *The prefrontal cortex: Anatomy, physiology, and neuropsychology of the frontal lobe* (3rd ed.). Philadelphia, PA: Lippincott-Raven.

Gilhooly, K. J., Logie, R. H., Wetherick, N. E., & Wynn, V. (1993). Working memory and strategies in syllogistic-reasoning tasks. *Memory & Cognition, 21*, 115-124.

Halford, G. S., Wilson, W. H., Guo, J., Gayler, R. W., Wiles, J., & Stewart, J. E. M. (1994). Connectionist implications for processing capacity limitations in analogies. In K. J. Holyoak & J. A. Barnden (Eds.), *Advances in connectionist and neural computation theory, Vol. 2: Analogical connections* (pp. 363-415). Norwood, NJ: Ablex Publishing.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: implications for comparative, developmental, and cognitive psychology. *Behavioral & Brain Sciences, 21*, 803-864.

Hummel, J. E., & Holyoak, K. J. (1992). Indirect analogical mapping, *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 516 - 521). Hillsdale, NJ: Erlbaum.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104*, 427-466.

Keane, M. T., Ledgeway, T., & Duff, S. (1994). Constraints on analogical mapping: A comparison of three models. *Cognitive Science, 18*, 387-438.

Klauer, K. C., Stegmaier, R., & Meiser, T. (1997). Working memory involvement in propositional and spatial reasoning. *Thinking & Reasoning, 3*, 9-47.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology, 25*, 431-467.

Norman, D. A., & Shallice, T. (1986). Attention to action: willed and automatic control of behaviour. In G. E. Schwartz & D. Shapiro (Eds.), *Consciousness and self-regulation* (Vol. 4). New York: Plenum Press.

Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities.* Hillsdale, NJ: Lawrence Erlbaum.

Sternberg, R. J., & Nigro, G. (1980). Developmental patterns in the solution of verbal analogies. *Child Development, 51*, 27-38.

Toms, M., Morris, N., & Ward, D. (1993). Working memory and conditional reasoning. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 46*, 679-699.

Waltz, J. A., Lau, A., Grewal, S. K., & Holyoak, K. J. (2000). The role of working memory in analogical mapping. *Memory & Cognition, 28*, 1205-1212.

Emotional Impact on Logic Deficits May Underlie Psychotic Delusions in Schizophrenia

Lilianne Rivka Mujica-Parodi (lrk9@columbia.edu)
Division of Medical Genetics, Laboratory of Clinical Neurobiology
New York State Psychiatric Institute—1051 Riverside Drive, Unit 6; New York, NY 10032 USA

Tsafrir Greenberg (tsafrirg@aol.com)
Division of Medical Genetics, Laboratory of Clinical Neurobiology
New York State Psychiatric Institute —1051 Riverside Drive, Unit 6; New York, NY 10032 USA

Robert M. Bilder (bilder@nki.rfmh.org)
Department of Medical Physics, Center for Advanced Brain Imaging—Nathan S. Kline Institute for Psychiatric Research
140 Old Orangeburg Road; Orangeburg, NY 10962 USA

Dolores Malaspina (dm9@columbia.edu)
Division of Medical Genetics, Laboratory of Clinical Neurobiology
New York State Psychiatric Institute —1051 Riverside Drive, Unit 6; New York, NY 10032 USA

Abstract

Psychotic delusions, defined as false immutable culturally discordant beliefs, constitute an endemic symptom among patients with schizophrenia. We examined whether the deficits in reasoning responsible for the formation and maintenance of delusions are a product of inappropriate discrimination between relevant and irrelevant data or rather the product of inappropriate processing from data that is appropriately relevant. We used a Logical Reasoning Task that presents items that test inferences and the choice of relevant information within two separate sections, one of which uses language designed to be affect-neutral, the other of which uses language with violent imagery that is designed to be affect-laden. Our preliminary results indicate that all patient groups show significant deficits on the types of reasoning that we measured, relative to healthy controls. While non-delusional patients also showed deficits in reasoning, delusional patients were unique in that their performance on reasoning tasks was most similar to that of healthy individuals *as long as the context was affect-neutral*. In the affect-laden contexts, however, delusional patients' reasoning significantly declined, while healthy individuals' performance was only mildly affected. We suggest that delusional patients may suffer from a vulnerability to arousal which causes them to commit types of reasoning errors similar in kind to those made by healthy controls under much more severe forms of stress.

Logic Deficits and Psychotic Delusions

A 36-yr-old man is arrested upon trying to enter the United Nations, arguing that he is Ambassador to Alpha Centauri; a 48-yr-old homeless woman believes that she is the wife of Thomas Aquinas; a 22-yr-old college student gradually becomes convinced that the CIA, FBI, and New York City Police Department have been following him since birth. All three of these individuals suffer from "psychotic delusions," the adoption of false, often bizarre, beliefs that are held in spite of ample evidence that contradicts the belief or greatly diminishes its likelihood. Psychotic delusions are characteristic of a number of mental and neurological illnesses, but are most common in schizophrenia and the manic phase of bipolar disorder. Delusional ideation produces much of the social alienation, lack of treatment compliance, and poor functioning associated with these diseases. Even so, the neurobiological and cognitive mechanisms responsible for the formation and maintenance of psychotic delusions are one of the least understood aspects of these illnesses.

In the absence of sensory distortion, it seems reasonable to assume that delusional patients have access to the same information about the world as everyone else. Yet if this is the case, we are left with two large puzzles. The first is why patients, who presumably start from the same premises as healthy individuals, manage to reach such radically different conclusions. The second is why patients' false conclusions revolve so consistently around certain themes, such as paranoia and grandiosity. Our study was designed to begin to address both questions.

Since logic was first developed to formalize rationality, it makes sense that it would serve as a useful tool in modeling aberrations of reason. If individuals infer conclusions from a set of premises by applying a pre-established category of rules of reasoning, then false conclusions may be arrived at by either starting from false premises or by invalid inferences. The vast majority of literature on schizophrenia and logic address the second of these possibilities, beginning with von Domarus' idea

that patients with schizophrenia consistently use a specific fallacious inference (Von Domarus, 1944). More modern studies have tested patients' abilities to use standard logical inferences (Ho, 1974; Kemp et al, 1997; Watson and Wold, 1981). Related to these are studies linking poor performance on the Wisconsin Card Sorting Task (a rule-generating task) with schizophrenia (Pantelis et al, 1999), and the correlation between delusion thought and a peculiar style of reasoning in which patients "jump to conclusions" (Huq, Garety, Hemsley, 1988). Results from these studies have been inconclusive. Though valuable as preliminary data, Mujica-Parodi, Malaspina, and Sackeim (2000) have argued that these studies are flawed because of the heterogeneity of patient samples, floor effects, and the conflating of different types of logic as if they reflected the same type of reasoning.

Surprisingly, little attention has been paid to the other possibility by which false conclusions may be reached: inappropriate choice of premises. This absence is all the more striking because modern empirical studies of normal cognition suggest a paradigm of reasoning, *mental models*, that is radically at odds with that presupposed by standard tests of logic (Johnson-Laird, 1995). The most obvious difference lies at the level of premises. Tests of deductive logic provide pieces of information that are explicitly described as the material from which conclusions ought to be derived. In the real world, however, our premises are seldom laid out so neatly before us. Instead, a large portion of our mental work must go towards discriminating between relevant and irrelevant information, choosing that from which we will later derive conclusions. Since available premise-groups are usually incomplete, most "conclusions" are actually closer to be being *hypotheses*, over-inclusive sets that are then restricted by confrontation with new evidence. Johnson-Laird's experiments suggest that the capacity for recognizing counter-examples to our provisionary models, and these models' subsequent revision, are just as critical in the formation of belief systems as the inferences that initially give rise to the models (Oakhill and Johnson-Laird, 1985). Since perhaps the most characteristic feature of delusions is not the strangeness of the conclusions reached, but of their perseverance in the face of systemic evidence to the contrary (Jones and Watson, 1997) one would expect the recognition and application of counter-examples to be a cognitive ability that is seriously impaired in patients with delusions. Yet it is a phenomenon that has hardly been studied in this population.

Our central hypothesis is that it is the failure to sort premises: distinguishing relevant from irrelevant information and, in particular, the recognition and application of counter-examples, that is responsible for delusions. We suggest that such a failure may be the result of a normal prioritization of neural resources during periods of emotional stress, inappropriately activated in patients. While it may be natural for healthy individuals to initially form false or partially false models, these models are normally revised in the face of contradictory evidence. In the presence of anxiety or fear, this self-correcting mechanism may be temporarily disabled in order to devote full mental resources to avoiding the cause of threat. Once the threat has passed, the mechanism would become functional once more and false beliefs would be revised. If, in the case of patients, the amygdala (and/or other medial temporal lobe structures responsible for perceiving threat) is hyperactive, then the same process may occur with far less provocation. Under the strain of a more or less continual state of emotional stress, the recognition of counter-examples may become disabled long enough to allow false beliefs to become entrenched. The advantage of grounding the creation of delusions in emotion, rather than the reverse, is that it provides an explanation for the relative thematic uniformity found between different patients' beliefs. The patient who experiences a state of fear or anxiety without a clear referent for it will presumably feel the need to explain that feeling; the explanatory structure that the patient creates, however, will very likely be flawed because of the mechanism hypothesized above.

The view of paranoid delusions as quasi-rational stories that are created to explain pre-existing emotional states is consistent with a wide variety of preliminary clinical evidence connecting psychosis and affect, including vague feelings of fear and anxiety reported during the prodromal (pre-symptomatic) period in schizophrenia (Henmi, 1993; Wiedemann *et al*, 1994) and correlations between levels of perceived stress and preoccupation with delusions (Myin-Germeys, 1999). It has long been known that there is a strong correlation between relapse in schizophrenia and even minor increases in stress (Doering et al, 1998). At a physiological level, schizophrenic patients' MRI's are characterized by lower hippocampal volumes than that of controls (Kegeles *et al*, 2000), these patients appear to have increased cortisol levels (a "stress hormone") (Goldman *et al*, 1993), and increased cortisol levels over sustained periods of time are known to be toxic to the hippocampus (Sapolsky *et al*, 1990). Other studies have linked prenatal damage to the limbic system with later development of schizophrenia (Lipska *et al*, 1993; Hanlon, Astur, and Sutherland, 1999). It may be that a vulnerable amygdala and/or medial temporal lobe abnormalities could be both responsible for the increased cortisol levels as well as for the common generalized feelings of anxiety and fear associated with the emergence of psychosis. It may also be the case that insult to the hippocampus and structures also involved with the circuit associated with arousal, such as cortical regions and the thalamus/hypothalamus, could be at least partly responsible for whatever defect in reasoning is responsible for the development of delusional beliefs. Most provocatively, this line of thought, if valid, could

shed light on the relationship between delusional ideation, clinically defined, and the anecdotal truism that even normally rational individuals become irrational when they are upset (or angry, or fearful, or euphoric, for that matter). It might also eventually provide a link between a particular symptom—delusions—and its neural mechanism among disorders, such as mania, psychotic depression, and schizophrenia, that are currently considered to be unrelated.

For our study, we tested the following hypotheses:

I. Patients with delusions are impaired in their ability to distinguish relevant from irrelevant premises in the formation and restriction of mental models.
II. Patients with delusions use logical inferences equivalently to healthy and patient controls.
III. All individuals (patients and controls) are impaired in their ability to distinguish relevant from irrelevant information when they are subject to emotional stress. We hypothesize that this task of sorting is more vulnerable to the effects of emotional stress than the task of logical inference-making.
IV. Patients are more vulnerable to stress, and therefore require less emotional provocation than do controls in order to reach the level at which this cognitive mechanism is impaired.

The Logical Reasoning Task

Dr. LR Mujica-Parodi and Dr. Harold Sackeim developed a *Logical Reasoning Task* (LRT) that avoids many of the problems present in previous attempts to test logic deficits in patients. Unlike previous tests, this task tests both logical inferences and the choice and evaluation of premises. The evaluation of premises sections are essentially an inverse of the inferences sections, in which the conclusion is given, and the subject is instructed either to identify information that would support the conclusion or to identify information that would contradict it. Items that are identical in form are presented within both affect-neutral and affect-laden contexts, counter-balanced for order. The affect-laden sections use threatening language in order to provoke a state of mild arousal, measured by visual analog scale and skin galvanic response. Responses are circled from a list of possible answers that are randomly spaced along the page. Extensive pilot testing ensured that LRT avoids floor effects by testing only inferences that most (60%-70%) healthy adults without any formal education in logic found to be intuitive and representative of "everyday reasoning." Most importantly, we established that both controls and patients are able to well-tolerate the task. The LRT is comprised of 60 items, 30 for the neutral section and 30 for the affect-laden section. For each section, 10 items test inferences, propositional premise choice, and class (quantifier) premise choice, respectively. The entire LRT takes approximately one hour to complete. Items on the LRT are scored in a manner that permits one to discriminate between errors of premise over-inclusion, errors of premise exclusion, and errors of contradiction. Subjects are able to choose more than one response. A score of 1 point is given for every item that includes the correct answer as long as there is no contradiction entailed by the subjects' responses. In Example 1 shown below (Table 1), the subject's choice of the second and third responses or the third and fourth responses would entail a contradiction. If the subject chooses more responses than are necessary, this is indicated by separate scoring for over-inclusion (with 1 additional point for each additional response). Similar scoring is done for under-inclusion. Separate scoring also records number of contradictions and choice of "not-enough-information" responses (which also indicate, with under-inclusion, degree of premise exclusion). Subjects are given five practice items before beginning the test, three of which have the correct answers marked, and two of which the subject completes.

Table 1: Examples from the Logical Reasoning Task

Example 1: Inferences, Affect-Neutral Condition
If John has missed the bus, then he will be late.
John has missed the bus.
What follows from this?
Nothing follows.
John will not be late
John will be late. ←
John has not missed the bus.
contradiction
contradiction

Example 2: Inferences, Affect-Laden Condition
If they are stabbing me, then they will kill me.
They are stabbing me.
What follows from this?
Nothing follows.
They will not kill me.
They will kill me. ←
They are not stabbing me.

Example 3: Premise Choice, Affect-Neutral Condition (Propositional logic, Counter-example)
John says that he will be late.
What information, together, makes you think that he is wrong?
John is late.
John has missed the bus.
Nothing makes me think that he is wrong.
John has not missed the bus. ←
Only if John is late, will he then miss the bus.
Only if John misses the bus, will he then be late. ←

Study Design

Subjects:
For these preliminary data we looked at responses from 28 patients, divided by cognitive symptom profiles, and 16 healthy controls. Of the patients, 10 had well-developed delusional systems, 5 were thought-disordered,

and 13 were neither delusional nor thought-disordered (with hallucinations as the primary symptom). Our completed data will contain 50 subjects in each of the 3 groups, as well as a group of healthy controls with high degrees of magical ideation. Diagnosis and symptom severity were determined using the Diagnostic Interview for Genetic Studies (DIGS). Patients were matched for symptom severity and medication status. For our final analysis, all three groups will be matched for Verbal IQ (using the Weschler Adult Intelligence Scales), age, and education. For our preliminary analysis, patients and controls were not matched due to the relatively small N, although ANCOVA's determined that covariates of education, gender, age, and education were not significant confounds. Both patient groups and controls were relatively well-educated, averaging two years of college, with no formal training in logic.

In our sample there were more male patients (21) than female patients (7), reflecting the general distribution in schizophrenia, and more female control subjects (10) than male control subjects (6). All subjects signed informed consent for this Institutional Review Board-approved study. Patients were recruited from the New York State Psychiatric Institute's Schizophrenia Research Unit, the Washington Heights Community Unit, and affiliated out-patient clinics. Controls were recruited from the local community, and were screened using the Psychosis Proneness Scales developed by Chapman and Chapman (Chapman et al, 1994).

Procedures:
Control subjects were screened using the Psychosis Proneness Scales. All subjects were administered the LRT and several sections of the WAIS (testing spatial inferences, abstraction without use of counter-examples, vocabulary, and working memory). Blind symptom-profiling for patients was performed post-hoc to avoid bias.

Results:
We performed ANOVA to determine differences between subject groups and test types (with post hoc t-tests), ANCOVA to screen confounds of age, sex, and education, and paired t-tests for individual subjects on affect neutral/laden condition to test the effects of arousal on performance. The results are summarized below in Table 2.

Differences between patient and control groups reached statistical significance for all types of reasoning (inferences: p=.005; premises(prop): p=.024; premises(class): .012). Performance was highest for healthy controls, followed by—in descending order—delusional patients, patients hallucinating only, and thought-disordered patients. This same pattern was present for all three types of reasoning. Healthy controls displayed a slight (non-significant) drop in performance when assessing relevance under mild arousal (premises(class): p=.083/df=15). Delusional patients displayed the same pattern, but in a significantly exaggerated form (inferences: p=.004/df=9; premises(class): p=.033/df=9) This suggests that delusional patients have a relatively intact ability to reason under neutral conditions, with a particular vulnerability toward emotional arousal. Thought-disordered patients displayed an inverse pattern, improving under emotional arousal (premises(class): p=.035/df=4), which perhaps reflects an inability to maintain adequate arousal under normal conditions. Healthy controls, delusional patients, and thought-disordered patients may be viewed as occupying different initial locations on an inverted "U," where—following the Yerkes-Dodson Law— performance initially improves with small degrees of arousal but suffers with increasing amounts of stress. In this case, stress-level is held constant, with different vulnerabilities to arousal accounting for the different locations on the curve. The types of errors made were also different for different groups. Delusional patients generally, and particularly under arousal, showed a tendency to shut out relevant information (p=.067/df=9), again an exaggeration of the healthy controls' response to stress (interestingly, patients who only hallucinated showed even more of an exaggeration in this regard). Thought-disordered patients, on the other hand, were more likely than other groups to assign inappropriate weight to irrelevant information (p=.05/f=2.3).

Table 2: 3 Types of Reasoning x 4 Subject Groups

	Inference	Prop	Class	
Del	0.54±22	0.31±28	0.46±.28	
Th Dis.	0.24±.18	0.11±.09	0.12±.08	**NEUTRAL**
Halluc	0.48±.20	0.29±.15	0.40±.27	**CONDITION**
HC	0.65±.19	0.45±.23	0.62±.28	
	Inference	Prop	Class	
Del	0.35±.24	0.24±.19	0.33±.25	
Th Dis.	0.26±.19	0.15±.10	0.20±.13	**AROUSED**
Halluc	0.41±.25	0.21±.13	0.28±.20	**CONDITION**
HC	0.67±.23	0.41±.24	0.52±.29	

Three important preliminary conclusions to be drawn from this data are that: a) all patients groups show significant deficits in both inferences and cognitive gating relative to controls; b) of all patients, delusional patients infer most similarly to healthy individuals, except in the presence of emotional material, which also seems to affect their reasoning most dramatically; and that c) the "irrationality" of delusional patients (which is affect-driven and shuts out relevant information) appears to be quite different from that of thought-disordered patients (which is affect-independent and reflects and inability to "screen-out" irrelevant information). This last point is of particular relevance because it suggests differences

between symptoms-types that may be relevant not only from the point of view of etiology, but of treatment (delusional patients may benefit from adjunctive benzodiazepines and/or cognitive-behavioral-therapy in a manner that thought-disordered patients may not, for instance). Interestingly, delusional patients' abnormal vulnerability to arousal, combined with healthy controls' less dramatic decline in cognitive gating under arousal, raise provocative questions regarding the degree to which delusional patients' reasoning may resemble normal controls' reasoning when normal controls are under more pronounced levels of stress.

Implications and Future Work

"Cognitive" versus "Sensory" Neural Gating

As shown above, our data suggests that all three patient groups had significant deficits in cognitive gating. In investigating the neurobiological roots to such a deficit, we have considered the possibility that deficits in cognitive gating may be rooted in a more fundamental "lower-level" deficit in "sensory-gating," whose deficits have been shown to be ubiquitous to schizophrenia. While "cognitive gating" is defined as the discrimination between relevant and irrelevant conceptual information, "sensory gating" can be similarly understood as the discrimination between relevant and irrelevant sensory data.

One explanation that could link cognitive and sensory gating is a "hierarchical" hypothesis for information processing, in which sensory data is sorted through a series of neural "filters," traveling first through a "coarse-grained" filter, then through filters that are progressively more "fine-grained" as attentional levels (and presumably levels of abstraction) are increased. This hypothesis is testable, for it predicts that poor performance on tests of higher-order filtering will also entail poor performance on tests of lower-order filtering. However, the converse is not true; performance on tests of lower-order filtering will not entail equivalent performance on tests of higher-order filtering (since information may be adequately filtered at a lower level before stumbling at a faulty higher-level filter). We would expect the brain activation associated with performance to behave equivalently. Thus we would predict that areas of the brain activated during higher-order filtering will entail (i.e., will correlate positively with) activation of areas of the brain associated with lower-order filtering (since all information that that passed through higher-order filters had to be processed through lower-order filters first). However, the converse will not be true; areas of the brain activated during lower-order filtering will not entail areas of the brain activated during higher-order filtering. M.-Marsel Mesulam [1998] proposes a similar hierarchical "critical gateway" model in which "lower" processing of sensory information contributes to "higher" processing of cognitive material along certain pre-established pathways. In Dr. Mesulam's model, the level of neurological impairment (i.e., whether related to "lower-order" global deficits like multimodal anomia or "higher-order" deficits like category-specific anomias) also results from the point in the hierarchy at which processing is disrupted. We are currently conducting a neuroimaging (fMRI) study looking at different information-gating processes (including the Logical Reasoning Task) operating at different "levels" to test this hypothesis.

While prominent, delusions are only one of many cognitive and perceptual symptoms of schizophrenia, which include thought-disorder and hallucinations. The elucidation of filter-level deficits, and their relationship to *specific* signs and symptoms of schizophrenia, is important in responding to a cogent criticism placed by Frith (1979), that the disabling of a sensory filtering mechanism in schizophrenia, if it exists, cannot be at a general level. This is because a generalized gating deficit would result in far more neurological disability than is actually clinically observed. However, the degree to which patients are disabled varies quite a bit, varying between extreme disorganization and intricately-constructed delusional belief systems. As Frith (1979) earlier has suggested, the precise symptomatology of the illness may be based on the *degree* to which gating is impaired. It may be that paranoid patients have relatively well-preserved lower-level filtering, while only higher-order filtering is impaired. Thought-disordered patients, on the other hand, may represent a more severe form of the illness in which lower as well as higher level gating is affected. Drawing these connections will be important not only theoretically, in establishing a unitary model of schizophrenia which accounts for varied symptoms, but also potentially in developing medications that are tailored to treating specific symptoms. If the mechanism associated with neural filtering does exist upon a continuum, ranging from the basic sensory gating involved in attention (at the lowest end), to the cognitive gating required to separate relevant from irrelevant information (at the highest end)—with each level dependent upon the one "beneath" it—then it may be the case that the stage at which filtering is impaired is responsible for the cognitive symptom picture (thought-disordered vs delusional, for instance) with which a schizophrenic patient presents. Higher-order filter deficits, with lower-order sensory gating that is still intact, may produce symptoms that look like belief systems that are flawed because they are constructed based on inappropriate choice of information (including inattention to counter-examples), but which remain fundamentally self-consistent. Patients whose filtering deficits have progressed to include *both* higher and lower-order gating, may produce a more general disorganization in which inferences, and therefore self-consistency, may no longer be possible. This picture is akin to Frith's model (Frith,

1979); he suggests that delusional ideation may exist as a less severe form of the disease, in which some aspects of cognition are still well-preserved. Based on the literature on sensory/sensorimotor gating and selective attention, we will be primarily looking at activation in the dorsolateral prefrontal cortex, hippocampus, striatum, and thalamus/hypothalamus.

Aknowledgments

Support for this research was provided by the National Alliance for Research on Schizophrenia and Depression, the Essel Foundation, the Frontier Fund for Scientific Research, and the National Institute for Mental Health.

References

Chapman L.J., Chapman J.P., Kwapil T.R., Eckblad M., & Zinser M.C. (1994). Putatively psychosis-prone subjects 10 years later. *Journal of Abnormal Psychology,* 103(2), 171-183.

Dawson M.E., Hazlett E.A., Filion D.L., Nuechterlein K.H., & Schell A.M. (1993). Attention and schizophrenia: impaired modulation of the startle reflex. *Journal of Abnormal Psychology,* 102(4), 633-641.

Doering S., Muller E., Kopcke W., Pietzcker A., Gaebel W., Linden M., Muller P., Muller-Spahn F., Tegeler J., & Schussler G. (1998). Predictors of relapse and rehospitalization in schizophrenia and schizoaffective disorder. *Schizophrenia Bulletin,* 24(1), 87-98.

Frith C.D. (1979). Consciousness, Information Processing and Schizophrenia. *British Journal of Psychiatry,* 134, 225-235.

Goldman M.B., Blake L., Marks R.C., Hedeker D., & Luchins D.J. (1993). Association of nonsuppression of cortisol on the DST with primary polydipsia in chronic schizophrenia, *American Journal of Psychiatry,*150(4), 653-655.

Hanlon F.M., Astur R.S., & Sutherland R.J.(1999). Changes in adult brain and behavior caused by neonatal limbic damage: a rat model with implications for understanding cognitive symptoms in schizophrenia (presented, *International Congress on Schizophrenia Research* 1999).

Henmi Y. (1993). Prodromal symptoms of relapse in schizophenic outpatients: retrospective and prospective study. *Japanese Journal of Pyschiatry and Neurology,* 47(4), 453-475.

Ho D.Y.F. (1974). Modern logic and schizophrenic thinking. *Genetic Psychology Monographs,* 89, 145-165.

Huq S.F., Garety P.A., & Hemsley D.R.(1988). Probabilistic judgements in deluded and non-deluded subjects. *Quarterly Journal of Experimental Psychology,* 40A, 801-812.

Johnson-Laird P.N. (1995). Mental models, deductive reasoning, and the brain. In Gazzaniga M.S. (Ed.),*The Cognitive Neurosciences*. Cambridge: MIT Press.

Jones E., & Watson J.P. (1997). Delusion, the overvalued idea and religious beliefs: a comparative analysis of their characteristics. *British Journal of Psychiatry,* 170, 381-386.

Kegeles L.S., Shungu D.C., Anjilvel S., Chan S., Ellis S.P., Xanthopoulos E., Malaspina D., Gorman J.M., Mann J.J., Laruelle M., & Kaufmann C.A. (2000).Hippocampal pathology in schizophrenia: magnetic resonance imaging and spectroscopy studies. *Psychiatry Research: Neuroimaging,* 98,163-175.

Kemp R., Chua S., McKenna P., & David A. (1997). Reasoning and delusions. *British Journal of Psychiatry,* 170, 398-405.

Lipska B.K., Jaskiw G.E., & Weinberger D.R. (1993). Postpubertal emergence of hyperresponsiveness to stress and to amphetamine after neonatal excitotoxic hippocampal damage: a potential animal model of schizophrenia. *Neuropsychopharmacology,* 9 (1), 67-75.

Mesulam M.M. (1998) From sensation to cognition. *Brain* 121, 1013-1052.

Mujica-Parodi L.R., Malaspina D., & Sackeim H.A. (2000). Logical processing, affect, and delusions in Schizophrenia: A New Cognitive Model, *Harvard Review of Psychiatry,* 8(2),73-83.

Myin-Germeys I. (1999). The contextual background of delusional experiences (presented International Congress on Schizophrenia Research)

Oakhill J.V., Johnson-Laird P.N.(1985). Rationality, memory and the search for counter-examples, *Cognition,* 20, 79-94.

Pantelis C., Barber F.Z., Barnes T.R., Nelson H.E., Owen A.M., & Robbins T.W. (1999). Comparison of set-shifting ability in patients with chronic schizophrenia and frontal lobe damage, *Schizophrenia Research,* 37(3), 251-270.

Sapolsky R.M., Uno H., Rebert C.S., & Finch C.E. (1990). Hippocampal damage associated with prolonged glucocorticoid exposure in primates. *Journal of Neuroscience,* 10(9), 2897-2902.

Von Domarus E. (1944). The specific laws of logic in schizophrenia. In J. Kasanin (Ed.), *Language and Thought in Schizophrenia*. Berkeley: University of California Press.

Watson C.G., & Wold J. (1981). Logical reasoning deficits in schizophrenia and brain damage. *Journal of Clinical Psychology,* 3(3), 466-471.

Wiedemann G.,Hahlweg K., Hank G. Feinstein E., Muller U., & Dose M. (1994). Detection of early warning signs in schizophrenic patients. *Nervenartzt* , 65(7), 438-43.

Interactions between Frequency Effects and Age of Acquisition Effects in a Connectionist Network

Paul W. Munro (munro@sis.pitt.edu)
School of Information Sciences
University of Pittsburgh
Pittsburgh, PA 15260 USA

Garrison Cottrell (gary@cs.ucsd.edu)
Department of Computer Science and Engineering 0114
University of California, San Diego
La Jolla, CA 92093-0114 USA

Abstract

The performance of a connectionist network, in which some resources are absent or damaged is examined as a function of various learning parameters. A learning environment is created by generating a set of random "prototypes" and clusters of exemplar vectors surrounding each prototype. An autoencoder is trained on the patterns. The robustness of each learned item is measured as a function of the time at which it was "acquired" by the network and its overall frequency in the environment. Both factors are shown to influence robustness under several learning conditions.

Introduction

For all their shortcomings, feed-forward network models of learning and memory share certain important features with their biological counterparts. Among these are the ability to gradually abstract statistical regularities from their environments by incorporating them into their connectivity structures and the feature generally known as "graceful degradation".

In this paper, the relationship between early learning (acquisition) and degradation of performance through loss of resources is examined in the context of small-scale simulations, in terms of frequency effects, age of acquisition (AoA) effects, prototype effects, and the insertion of noise into the neural network.

The relative influence of AoA compared to frequency on word naming tasks has been argued among cognitive psychologists and linguists for several years now (Brown & Watson, 1987; Morrison et al., 1992; Gerhard & Barry, 1998). Of course, teasing apart the influences of AoA and frequency is confounded by the strong correlation between them. AoA effects have also been reported in other domains, such as object identification and face recognition (Moore & Valentine, 1999). The effects of AoA and frequency on pattern error have been analyzed by Smith, Cottrell, and Anderson (2001). Here, we look at pattern performance in the face of damage to the network, simulating neuronal failure as could occur with aging or trauma.

The robustness of network performance to hidden unit damage has been shown to improve for networks trained with noise among the hidden units (Judd & Munro, 1993). In some cases, this kind of noise has been shown to improve the generalization properties of a network (Clay & Sequin, 1990). Functionally, the hidden representations of the training items settle to states that are further apart in terms of a Euclidean measure.

In this paper, we examine the following three hypotheses:

1. The robustness of an item under loss of network computational resources (analogous to the loss of neurons in humans) is related both to the time at which that item was "acquired", and to the average frequency of the item in the network's experience.

2. Prototypical items are more robust than exemplars, even if they are never explicitly presented to the network, since they share features with populations of exemplars, and thus have high "effective frequencies" in the environment.

3. Early explicit learning of prototypes can result in a more robust set of internal exemplar representations.

Methodology

The training set

A two-step process is used to generate a structured set of bit strings of length L. First, a set of N prototype strings is produced by generating 0 and 1 values

independently with probability 0.5 for each bit having a value 1. In the second step, a set of n_i exemplar strings are generated from the i^{th} prototype P_i by "flipping" bits with a low probability. The result is a set of N pattern "clusters" (see Figure 1). While the network is trained on the exemplar patterns only, the network performance is measured for both the exemplars and the prototypes. In this study, L=100, N=10, and n_i=10, (i=1..10).

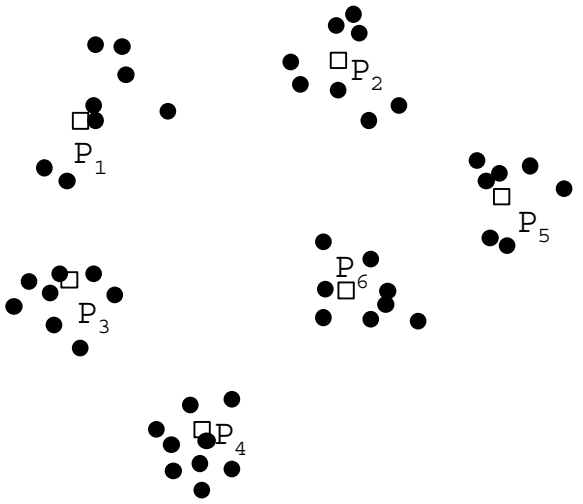


Figure 1. Schematic view of the training patterns. Prototypes (open squares) are randomly generated binary strings. A set of exemplars (filled circles) is generated in the neighborhood of prototype.

In order to analyze the role of frequency, items were selected from the training set according to a ramp distribution; that is, the selection probabilities for the 100 exemplars ranged from approximately 0.0002 to 0.0200 linearly. The probabilities were assigned such that the cluster probabilities also followed a ramp distribution. In other words, the items were ordered according to the parent prototypes and the probability of selecting the k^{th} item was proportional to k. This way, the clusters were also ordered such that the probability of selecting an item from the j^{th} cluster was proportional to j.

Network Architecture

Networks trained by backpropagation to reconstruct their input pattern at the output layer (autoencoders) with a single hidden layer of 40 units are trained using backprop. In some trials, the responses of the hidden units are randomly perturbed to analyze the effect of network noise. An output unit's response is deemed "correct" if it differs from the target by less than a predetermined tolerance level δ. Performance is measured in terms of the number of correct output units. If the network responds with a sufficient number of correct output units to an input pattern, that pattern has been acquired by the network. The point in training at which a pattern is first acquired is called its age of acquisition (AoA). Preliminary studies have shown that in some cases a pattern may briefly be "forgotten" soon after its initial acquisition. In such instances, the forgotten pattern is promptly reacquired; thus, the AoA is defined as the time the pattern is first acquired.

Performance Analysis

After training, the network's response to each training pattern was tested under various damage levels. Damage was implemented by only allowing the output of k of the H hidden units to stimulate the output layer, where k is varied from 1 to H. The minimum number of hidden units required to reconstruct the input pattern (to within a specified degree of tolerance) is recorded as a measure of the pattern's robustness in the network. In some cases, patterns were "forgotten" after initial acquisition. In most such cases, the pattern was reacquired, but not always.

Experimental Conditions

In all the experiments, the acquisition criterion is that 95 out of 100 units should be within 0.2 of their target value (0 or 1). The total training time is either 50000 or 100000 pattern presentations, depending on the condition. Thus, with the ramp distribution, the number of presentations of each individual pattern varies from about 10 to about 2000.

Control Condition (CC) In the control condition, the network is trained with just the 100 exemplars for a period of 100000 pattern presentations.

Head Start Condition (HC) Here, the training set consists a subset of only 10 patterns (one from each cluster) of the full set of 100 exemplars for the first 10000 time steps. This is done to guarantee very low AoAs on some patterns. The training set is expanded to the full set, including the initial subset, for 90000 more presentations. Ellis & Lambon-Ralph (2000) found strong AoA effects in a staged learning condition of this kind.

Noisy Condition (NC) This condition is the same as the previous condition (HC) with "Boolean" noise injected into the hidden layer during the early phase. Here, the activity levels of a small number of hidden units are multiplied by −1. This manipulation is predicted to increase the overall robustness of the full training set.

Prototype Condition (PC) In this variation of HC, the network is trained on only the protoypes during the early phase with no injected noise. Note that prototypes

are never explicitly presented in the previous three conditions.

Results

All conditions show a strong dependence of AoA on frequency. In general, prototype patterns are acquired earlier than exemplar patterns, even if they are not explicitly presented, with the AoA of the prototypes dependent on average frequency of the corresponding exemplars.

Control Condition (CC)

Over the course of 100000 exemplar pattern presentations, 92 of the 100 exemplars were acquired by the network. The eight nonacquired exemplars were all among the 11 least frequent. Of the 10 prototypes, one was not acquired, and eight were acquired in the first 10000 iterations. A scatterplot of AoA vs frequency follows a hyperbolic trend (Fig 2, top). This observation prompted a second scatterplot (Fig 2, bottom), in which AoA is examined vs. $freq^{-1}$. Regression on these data indicates the product of AoA and frequency is about 190 (zero intercept assumed).

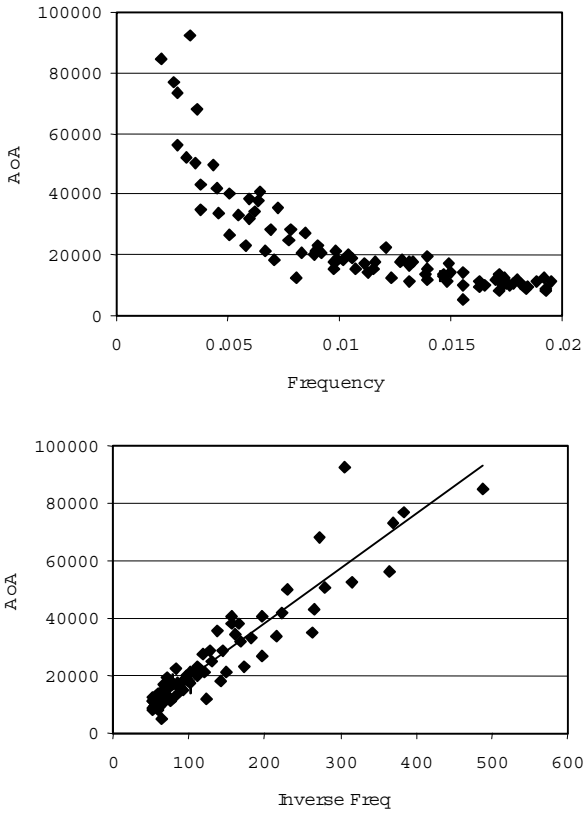


Figure 2: AoA vs Frequency (top) and AoA vs. $Freq^{-1}$ (bottom). The random selection of stimuli in the simulation follows a ramp distribution to give a wide range of frequencies.

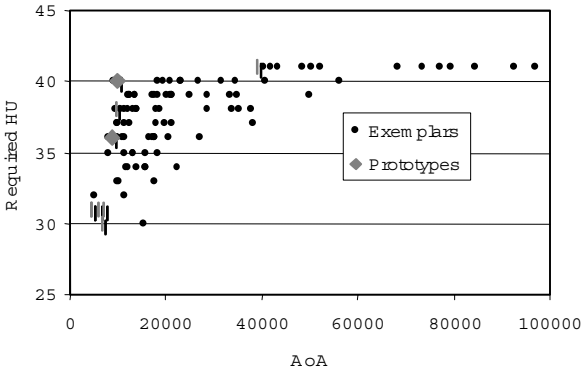


Figure 3. The number of hidden units required to reconstruct the input as a function the AoA. A value of 41 indicates that when the simulation halted, the pattern could not be reconstructed with all 40 hidden units.

The fragility of each item, as measured by the number of hidden units required to reconstruct the pattern tends to be higher for the patterns with later AoA (i.e., earlier patterns are more robust). This is true for both the exemplars and the prototypes (Fig 3). Similarly, items that are more frequent tend to be more robust (Fig 4).

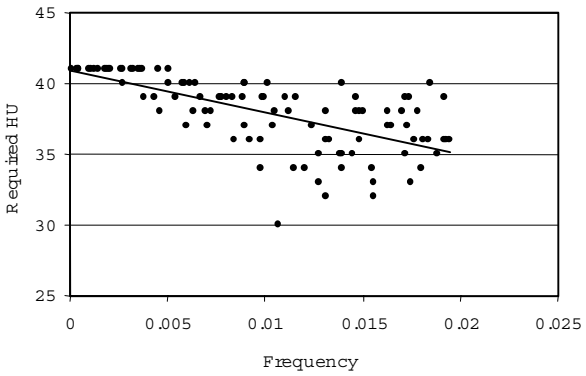


Figure 4. The required number of hidden units vs. frequency. The trendline shows that more frequent items tend to be more robust.

Regression against both variables indicates that the influence of AoA ($p = 0.01139$) is stronger than frequency ($p = 0.03271$) by a factor of almost three.

Head Start Condition (HC)

Here, the first 5000 iterations use only a subset of 10 items (one exemplar from each prototype's "cluster") for training. The network is then exposed to the entire set of 100 exemplars for 45000 subsequent learning trials. Selection of patterns during early exposure also follows a ramp distribution, giving a variety of frequencies within this set.

Early Items. Nine of the 10 items presented alone for the first 5000 time steps are learned before presentation 2000. Four of them are acquired before the earliest prototype (1000 iterations). The least frequent item in this set was never learned. As in CC, AoA and frequency are highly correlated.

Prototypes. The mean AoA for prototypes under HC (12907) is later than it is under CC (10568) and the average prototype is slightly less robust under HC (35.75 HU) than under CC (34.22 HU).

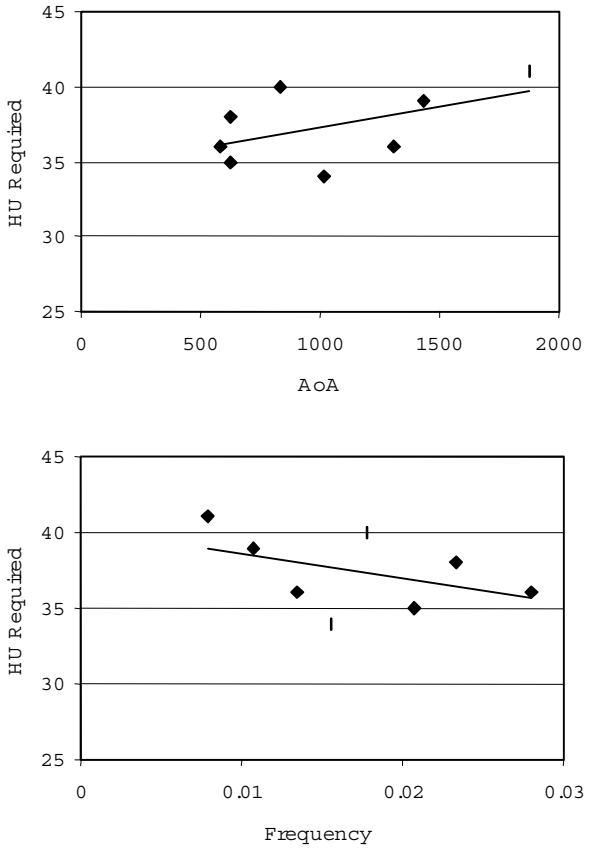


Figure 5. The dependencies of robustness on AoA (top) and frequency (bottom) under HC.

Prototype Condition (PC)

This condition is like HC, except that the ten patterns presented in the early phase are the prototypes of the later patterns. No significant differences in the effects on robustness or AoA were observed in the PC relative to HC.

Noisy Condition (NC)

As in the case of PC, this condition produced mainly negative results. No significant effect of the noise was noticed on the acquisition or robustness of the exemplars. The main observed effect of noise is that the prototypes are acquired much faster. However, the network does not maintain the ability to reconstruct prototypes from the low frequency clusters. Nevertheless, those prototypes that are maintained can withstand more damage to the network.

The bar graphs in Figure 6 display the AoA and robustness (HU required) for the prototype patterns, such that they can be compared with corresponding values in the control condition (black bars=NC, striped bars=CC).

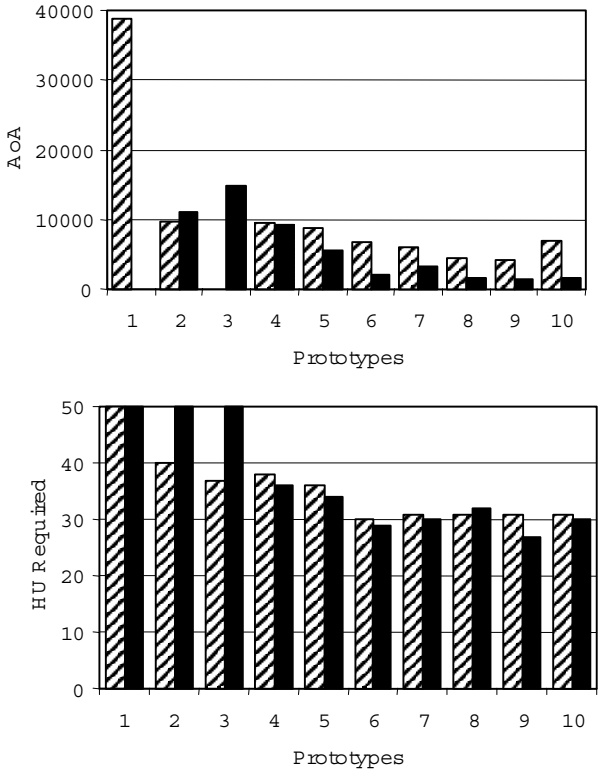


Figure 6. Distributions of AoA (top) and the required number of hidden units (bottom) are displayed for the control condition (striped bars) and NC (dark bars). The 10 items are ranked from lowest (approx. 0.018) to highest frequency (approx. 0.18). The network was never able to reconstruct the lowest frequency prototype (#1), hence there is no bar for this condition. The maximum value for the lower bar graph is the total number of hidden units, 40. A value of 50 means that the network could not reconstruct those prototypes at the end of the simulation.

Conclusions

As a preamble to the data analysis, the relationship between AoA and frequency was examined. These variables were found to be strongly related by a function of the form $a=k/f$, where a is the AoA, f is the frequency, and k is a constant (refer to Fig 2). Even

though this did not bear directly on the hypotheses, it may be the strongest result of this paper!

Our results support the first two hypotheses. The first hypothesis, that both frequency and AoA influence robustness of a learned item is evident from the simulations. Bivariate regression of the robustness variable (HU required) against the two independent variables gave fits that were not very tight (i.e., the p values were too high for the results to be considered significant). Nevertheless, the value corresponding to AoA was consistently lower than that for frequency, indicating a stronger dependence of robustness on AoA.

The second hypothesis, that prototypes are more robust than exemplars was supported by the simulations. The effect is as strong as expected by the measure used here: under CC, prototypes require an average of 34.3 HU, while exemplars require 36.3 HU. Note that this may simply by a byproduct of the AoA effect, since prototypes are acquired much earlier than exemplars. Frequency also plays a role. Even when the prototypes are not explicitly presented, and thus have no frequency per se, the exemplars may be considered distorted versions of the prototypes. Hence, each prototype has an "effective frequency" that depends on the total frequency of its supporting exemplars weighted by the exemplar-prototype distances.

Our simulations did not support the third hypothesis, that early explicit prototype training would result in representations that are more robust. While no such effect has yet been observed, it remains as a subject for future investigation.

Discussion

The issues investigated in this study are the first steps into the exploration of a broader question: How does the adult cognitive structure ultimately depend on the initial stages of learning? This question is quite similar to the age-old debate of nature vs. nurture. Here the issue is whether some potential for later cognitive capabilities is dependent, not on innate factors, but on the content of early experience and the biological mechanisms at work.

The process of acquisition of information, the sequence in which items are presented to the learner, as well as the internal parameters of the learner, may play a determining role in the adult conceptual architecture. It may be that the representations of concepts acquired in childhood, and the associations formed among them construct a foundation on which later concepts are built. Hence, the soundness of this foundation may determine the ultimate robustness of the adult.

Certainly, the importance of early learning on cognitive development has been acknowledged (for example, Catherwood, 1999). In the present work, we have begun to examine this within the connectionist framework, whereby adult cognitive performance might be linked to the statistics of the learning environment in early childhood.

.

Acknowledgments

We would like to thank the members of the GURU group at UCSD for valuable discussions. Paul Munro gratefully acknowledges the hospitality of the group and its leader, coauthor Garrison Cottrell, during a sabbatical leave spent at UCSD during the fall quarter of 2000.

References

Brown, G. & Watson, F. (1987) First in, first out: Word learning age and spoken word frequency as predictors of word familiarity and naming latency. Memory & Cognition, 15:208-216

Catherwood, D. (1999) New Views on the Young Brain: offerings from developmental psychology to early childhood education. Contemporary Issues in Early Childhood, 1:23-35.

Clay, R. & Séquin, C. (1992) Fault tolerance training improves generalization and robustness. In: Proceedings of the International Joint Conference on Neural Networks. IEEE/INNS: Baltimore MD.

Ellis, A. W., & Lambon-Ralph, M. A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. Journal of Experimental Psychology: Learning, Memory, & Cognition, 26(5), 1103-1123.

Ghiselli-Crippa, T. & Munro, P. (1994) Emergence of global structure from local associations. In: Advances in Neural Information Processing Systems 6, J.D. Cowan, G. Tesauro, J. Alspector, eds. San Mateo, CA: Morgan Kaufmann.

Judd, S. & Munro, P. (1993) Nets with unreliable hidden nodes learn error-correcting codes. In: Giles, C.L., Hanson, S.J., Cowan, J.D., (eds.) Advances in Neural Information Processing Systems 5. San Mateo, CA: Morgan Kaufmann Publishers.

Moore, V. & Valentine, T. (1999) The effects of age-of-acquisition in processing famous faces and names: Exploring the locus and proposing a mechanism. Proceedings of the Twenty-First Annual Meeting of the Cognitive Science Society. Mahwah NJ: Erlbaum.

Smith, M., Cottrell, G., and Anderson K. (2001) The early word catches the weights. To appear in: Advances in Neural Information Processing Systems 12 MIT Press, Cambridge, MA.

Modality preference and its change in the course of development

Amanda C. Napolitano (napolitano.7@osu.edu)
Department of Psychology
Ohio State University
48 Townshend Hall, 1885 Neil Avenue
Columbus, OH 43210, USA

Vladimir M. Sloutsky (sloutsky.1@osu.edu)
Center for Cognitive Science & School of Teaching & Learning
Ohio State University
21 Page Hall,1810 College Road
Columbus, OH 43210, USA

Sarah T. Boysen (boysen.1@osu.edu)
Department of Psychology
Ohio State University
48 Townshend Hall, 1885 Neil Avenue
Columbus, OH 43210, USA

Abstract

The current study examines the modality preference and its change in the course of development. Based on findings from previous research (Balaban & Waxman, 1997; Roberts, 1995; Sloutsky & Lo, 1999), it was expected that the auditory modality would be privileged at a very young age. In the experiment, participants after being trained to select a target image and sound combination were required to select a new combination where the target image was paired with a new sound and the target sound was paired with a new image. It was argued that the selected set would be indicative of whether the image or sound is more salient. Results indicate that the auditory modality was more salient than visual for 4-year-olds, whereas the visual modality was more salient for 5-year-olds and adults.

Introduction

It is well known that linguistic labels play an important role in induction. For example, in Gelman & Markman's (1986) forced-choice task, young children were presented with pictures of a blackbird (Target), flamingo (Test 1), and bat (Test 2) that was perceptually similar to the blackbird. Both the blackbird and flamingo were referred to as "birds," and the task was to induce a property (e.g., "feeds its young with mashed food" vs. "feeds its young with hard food") from one of the Test stimuli to the Target. Results indicated that young children reliably induced biological properties from one bird to another bird, even when both birds were perceptually dissimilar. Other researchers demonstrated large, albeit quantitative, effects of labels using a variety of stimuli and tasks (Sloutsky & Lo, 1999; 2000; Sloutsky, Lo, & Fisher, in press). There are also findings that in categorization tasks linguistic labels supported taxonomical grouping of objects in infants (Balaban & Waxman, 1997; Waxman & Markow, 1995) and in young children (Markman, 1989).

These findings have generated three primary explanations: semantic, phonetic, and general-auditory. The two former explanations argue for specific linguistic effects of labels, whereas the third argues for more general auditory effects.

The semantic explanation (Gelman & Markman, 1986; Markman, 1989) has focused primarily on explaining the effects of labels on induction in young preschool-age children. It has been argued that linguistic labels presented as count nouns denote categories, and, because members of the same category are likely to share many nonobvious properties (two cows are more likely to have similar insides than a cow and a pig) shared linguistic labels support induction.

The phonetic explanation (Balaban & Waxman, 1997) has focused on explaining categorization in young infants. According to this position, young infants are especially attentive to the prosodic components of human speech that distinguish it from other sounds (Balaban & Waxman, 1997, Experiment 3). As a result, when presented with pictures accompanied by auditorily introduced labels, infants attended to linguistic labels and to visual features that correlated with linguistic labels.

Finally, the general auditory explanation suggests that, at least for infants and very young children, the effects of labels may stem from the modality of input (Sloutsky & Lo, 1999). In particular, it is possible that the powerful effects linguistic labels have on categorization and induction are due to the fact that auditory input has a privileged processing and attentional status in younger humans (cf. Roberts, 1995; Roberts & Jacob, 1991). If this is the case, stimuli

presented auditorily should have larger attentional weights than stimuli presented in another modality.

Why would auditorily presented stimuli weigh more for younger children than visual stimuli? One possible explanation is that the auditory system matures earlier than the visual system: in particular, the auditory system starts functioning during the last trimester of gestation (Birnholz & Benaceraff, 1983; see also Jusczyk, 1998, for a review), whereas the visual system does not start functioning until after birth. As a result, even though the neural bases of visual perception are fully developed at quite a young age (e.g., Aslin & Smith, 1988), auditory stimuli may still have a privileged processing status for younger children, thus resulting in larger weights of auditory stimuli. This privileged status of the auditory modality may be functionally important for language acquisition, and, in this case the advantage may start decreasing when the child has (in principle) acquired the task of acquiring language.

The goal of current research is to test this general auditory explanation. Note that support of the general auditory explanation does not rule out the semantic and the phonetic explanations. Because the three explanations are not mutually exclusive, it is possible that linguistic labels may have semantic and phonetic effects above and beyond the general auditory effect.

The overall experimental approach was as follows. Participants were presented with two stimulus sets each consisting of an auditory and a visual component and were trained to consistently select one set over the other. When training was accomplished, they moved to a test phase, in which the trained set was split, such that the visual stimulus of the trained set was paired with a novel auditory stimulus, whereas the trained auditory stimulus was paired with a novel visual stimulus. The participants were asked which of the two was the trained set. It was argued that if participants put more weight on the visual stimulus, they should select the first set, whereas if they put more weight on the auditory stimulus they should select the second one.

Method

Participants

A total of 39 children and undergraduate students participated in the experiment. Participants represented three age groups with 13 participants in each group: (1) 48-month-olds to 57-month-olds, (2) 58-month-olds to 65-month-olds, and (3) undergraduate students at The Ohio State University. The second group was added when the experiment was under way. This was done to provide more dense developmental observations. Children participants were recruited from local childcare centers in the Columbus, Ohio area. Undergraduate students participated as part as part of an introductory psychology course.

Materials

Materials consisted of 24 stimulus sets. Each set was comprised of a visual and an auditory stimulus. The visual stimuli were digitized photographic landscape images. Each image consisted primarily of a different type of green foliage. Images were 4 inches by 4 inches in size. The auditory stimuli were computer generated patterns, each consisting of three unique simple tones. Simple tones varied on flavor (sine, triangle, or sawtooth) and frequency (between 1 Hz and 100 Hz) components. Each simple tone was .3 seconds in duration and was separated by .05 seconds of silence, for total pattern duration of 1 second.

Diagnostics were run to insure the auditory and visual stimuli had high discriminability. This was accomplished using a same-different task. In the task participants were presented with one stimulus for a duration of one second, followed by the presentation of a second stimulus for a duration of one second, and then asked whether the two stimuli were the same. The participants were able to discriminate between all stimuli on over 95 percent of the trials.

Design and Procedure

The experiment included six blocks, each consisting of 8 training trials (a training session) and 6 test trials (a testing session). In each block, 4 out of the 24 stimulus sets were used. Two of these four sets were used in the training session, and the other two were used in the testing session. Children participants were tested in a quiet room within their daycare center. Small toys were used as rewards for participation. Undergraduate participants were tested in a lab on campus. A laptop computer controlled presentation of stimulus sets and recorded all responses. Participants entered the room and sat in a chair in front of laptop. They were told that they would play a game (references to toys were omitted for undergraduate participants), in which they should find the location of a prize. They were then presented on-screen with two stimulus sets each consisting of a visual component (V_1 vs. V_2) and an auditory component (A_1 vs. A_2).

Stimuli were presented in the following manner. First, V_1 and A_1 were presented simultaneously on one side of the screen, followed by the presentation V_2 and A_2 on the other side of the screen. Each image's presentation matched the duration of its sound, and was replaced by a white circle icon at the end of each set's presentation. In short, the child was presented with two stimuli sets V_1A_1 and V_2A_2 and the task was to identify the stimulus set, under which the prize is hidden. The goal of training was to teach the child to consistently

select a particular stimulus set, and, therefore, on each trial the child was provided with yes/no feedback. The positions of each of the two stimulus sets were counterbalanced across the 8 training trials. Participants making correct selections in the final four trials moved into the test session.

The test session followed immediately after the training session, during which participants were presented with two novel stimulus sets: set one (V_1A_{new}) matched the training target's visual component, but had a novel auditory component, whereas set two ($V_{new}A_1$) had a novel visual component, but matched the training target's auditory component. The participants were asked again to identify the one where a prize was hidden. When the participant's selection was made, the experimenter pressed the keyboard key corresponding to the selection, without giving feedback to the participant. The overall structure of training and testing trials is presented in Table 1.

Table 1: The overall structure of training and testing trials.

Training Trial		Testing Trial	
V_1A_1 (Target)	V_2A_2 (Distracter)	V_1A_{new}	$V_{new}A_1$

Results and Discussion

In this section, we analyze participant's choices in the testing phase. Recall that stimulus sets were arranged such that participants could rely either on the visual components of the learned stimulus set (V_1) or on the auditory component (A_1). We first compare overall means of visual and auditory responses across the three age groups. We then report provide a more detailed analysis of participants' performance. In particular, we compare the number of blocks where participants were above chance selecting either the familiar visual or auditory component. We also analyze individual patterns of responses, comparing the number of participants consistently exhibiting auditory responding with those consistently exhibiting visual responding. Note that 48-month-olds to 57-month-olds successfully accomplished 61 out of 78 training sessions (78%), 58-month-olds to 65-month-olds successfully accomplished 65 out of 78 training sessions (84%), and adults successfully accomplished all 78 training sessions. There were also 7 children in the youngest group and 2 children in the older group who did not pass a single training session. These children were eliminated from the analyses and they are not a part of 39 participants whose data are reported here.

Overall means for auditory-based responding were subjected to a one-way ANOVA with age as a factor, followed up by post-hoc Tukey tests. The analyses indicated that these means (65% vs. 22% vs. 2%) differed significantly across the three age groups, F (2, 36) = 19.9, $p < .0001$, and post-hoc Tukey test confirmed that there were significant differences among the groups.

To analyze participants' performance in test sessions, we calculated the number of sessions with above-chance reliance on auditory stimuli, above chance reliance on visual stimuli, and chance performance. Performance was considered above chance if the same choice was made on 5 out of 6 trials (Binomial Test, $p = .09$), otherwise it was considered at chance. Results indicate that in the group of 48-57-month-olds only 9 out of 61 session were at chance, and in the group of 58-65-month-olds 10 out 65 were at chance. All other sessions were above chance. In the group of undergraduate students all test sessions were above chance. Percentages of sessions with above-chance performance by age group and stimulus modality are presented in Figure 1.

Percentages of sessions with above change auditory and responses were subjected to two separate one-way ANOVA with age as a factor. There were significant differences across age groups, both Fs (2, 36) >19, ps < .0001. The post-hoc Tukey tests pointed to the following order differences: 48-57-month-olds were more likely to rely on auditory stimuli and less likely to rely on visual stimuli than 58-65-month-olds or undergraduate students.

Across age groups, there emerged three distinct patterns of responses: (1) participants who were above chance in relying on auditory stimuli (auditory responders); (2) participants who were above chance in relying on visual stimuli (visual responders); and (3) participants who were at chance (mixed responders). Percentages of responders' types across age groups are presented in Table 2.

Table 2. Percentages of responder types by age group.

	Responder Type		
Age Group	Visual	Auditory	Mixed
48-57-month-olds	15.38	61.53	20.07
58-65-month-olds	76.92	15.38	7.69
19-year-olds	100.00	0	0

Numbers of auditory and visual responders in each age were subjected to a chi-square analysis. The analysis pointed to significant differences among the groups, both χ^2s (2, N =39) > 13.4, $p < .001$. The analysis of standardized residuals indicated that 48-57-month-olds were more likely to exhibit an auditory-based pattern than 58-65-month-olds or 19-year-olds, whereas 58-65-month-olds or 19-year-olds were more likely to exhibit a visual-based pattern of responses (all zs > 3.1, ps < .001).

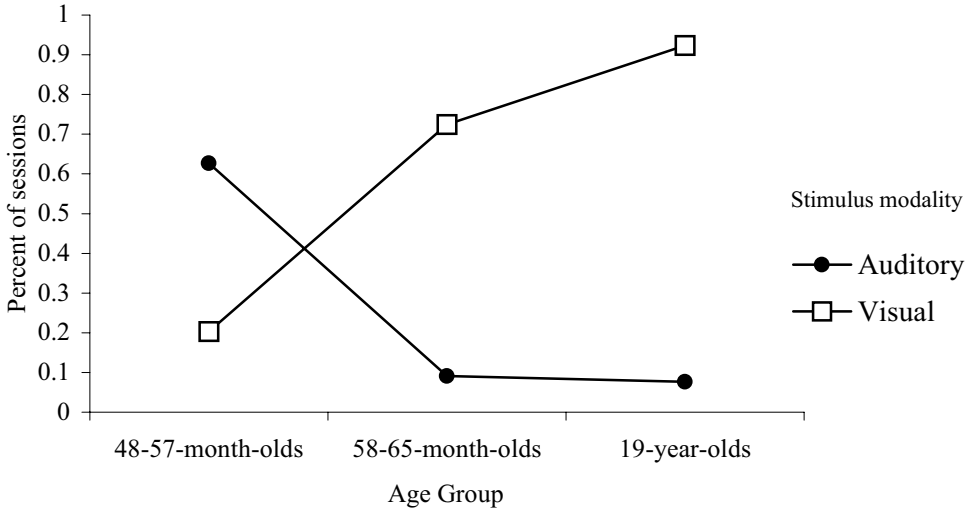


Figure 1. Percent of test seccions at above-chance performance by age group and stimulus modality

In short, the reported data indicate that for 48-57-month-olds auditorily presented stimuli weighed more than visually presented stimuli. These data support the auditory explanation predicting larger weights of auditory stimuli weights for younger children, but not for older children or adults. Of course, the support of the general auditory explanation does not rule out either the phonetic or the semantic explanations. These explanations will be further examined in our future research.

There is also a possible alternative explanation for the significant differences between 48-57-month-olds and the two other groups. First, the younger children's selections may be due to differential complexity of the stimuli. Each auditory stimulus contains far fewer features as compared with the visual stimuli, and, for this reason, the auditory stimuli are simpler to encode and process. Therefore, it is possible that children in the youngest group prefer simpler stimuli that happened to be auditorily presented, rather than prefer auditory stimuli per se. Of course, this explanation would have a hard time explaining the sudden shift from the larger weight of auditory stimuli observed in 48-57-month-olds to the larger weight of visual stimuli observed in 58-65-month-olds, however our current hypothesis also does not have an acceptable explanation for the shift. As mentioned above, there are two possible reasons for audition having priority at a young age. The auditory modality is functional before birth, and it is clear that at birth it is the dominant modality. Perhaps, it is not until the late 4's that the visual modality gains its privileged status. The privileged status of audition may also be related to language acquisition. It is possible, that the auditory modality is privileged during the period where a child's vocabulary acquisition is at its highest. In both cases, our experiments are likely to capture the very end of the period when auditory modality is privileged and the transition to the privileged status of the visual modality. To clarify this issue additional experiments with young infants are needed.

In future research, we plan to examine the above mentioned alternative explanations suggesting that the observed results stem from different computational complexity of stimuli, with impoverished auditory patterns being more simple than more rich and elaborated visual scenes. The next phase of the study will be to reverse the complexity of the stimuli and present participants with impoverished visual stimuli and complex elaborated auditory stimuli. The visual stimuli will be computer-generated two-dimensional geometric figures. The auditory stimuli will be compressed pieces of Celtic music. If the 48-57-month-old group continues to make selections that favor audition, this will provide evidence for the privileged status of auditory processing for young children. If the 48-57-month-olds' selections favor visual stimuli, this will instead provide evidence that young children use stimulus information that is easier to process.

Following this second experiment, it will be useful to examine whether introducing human speech has any effect on the percentage of auditory selections. In this experiment, sounds will be comprised of three phonemes (e.g. "ba te do"). The results of this

experimentation, considered with the results from experiment one, should elicit the attentional weights given to language-rich and auditory stimuli in general.

Finally, it will be necessary to test younger children and infants to determine if preference for auditory stimuli decreases monotonically with age. It is possible that the relation is an inverted U-type, and the auditory modality only becomes dominant during the "explosion" period of language development.

While these alternatives will be tested in the future, research presented here indicates that under these conditions, auditory stimuli weigh more than visual stimuli for 48-57-month-olds, whereas visual stimuli weigh more than auditory for older children and adults. These results support the hypothesis that for young children auditory input may have privileged processing status.

Acknowledgments

This research has been supported by a grant from the National Science Foundation to the second author.

References

Aslin, R., & Smith, L. (1988). Perceptual Development. *Annual Review of Psychology*, 39, 435-474.

Balaban, M. T. & Waxman, S. R. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child Psychology, 64*, 3-26.

Birnholz, J. C., & Benaceraff, B. B. (1983). The development of human fetal hearing. *Science*, *222*, 516-518.

Gelman, S. A., & Markman, E. (1986). Categories and induction in young children. *Cognition*, *23*, 183-209.

Jusczyk, P. W. (1998). *The discovery of spoken language*. Cambridge, MA: MIT Press.

Markman, E. (1989). *Categorization and naming in children*. Cambridge, MA: MIT Press.

Roberts, K (1995). Categorical responding in 15-month-olds: Influence of the noun-category bias and the covariation between visual fixation and auditory input. *Cognitive Development, 10*, 21-41.

Roberts, K. & Jacob, M. (1991). Linguistic versus attentional influences on nonlinguistic categorization in 15-month-old infants. *Cognitive Development, 6,* 355-375.

Sloutsky, V. M. & Lo, Y. (1999). How much does a shared name make things similar? Part 1: Linguistic labels and the development of similarity judgment. *Developmental Psychology*.

Sloutsky, V. M., & Lo, Y.-F. (2000). Linguistic labels and the development of inductive Inference. *Proceedings of the XXII Annual Conference of the Cognitive Science Society* (pp. 469-474). Mahwah, NJ: Erlbaum.

Sloutsky, V. M., Lo, Y.-F., & Fisher, A. V. (In Press). How much does a shared name make things similar? Linguistic labels and the development of inductive inference. *Child Development*.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, *29*, 557-302.

Clustering Using the Contrast Model

Daniel J. Navarro and Michael D. Lee
daniel.navarro,michael.lee @psychology.adelaide.edu.au
Department of Psychology, University of Adelaide
South Australia, 5005, AUSTRALIA

Abstract

An algorithm is developed for generating featural representations from similarity data using Tversky's (1977) Contrast Model. Unlike previous additive clustering approaches, the algorithm fits a representational model that allows for stimulus similarity to be measured in terms of both common and distinctive features. The important issue of striking an appropriate balance between data fit and representational complexity is addressed through the use of the Geometric Complexity Criterion to guide model selection. The ability of the algorithm to recover known featural representations from noisy data is tested, and it is also applied to real data measuring the similarity of kinship terms.

Introduction

Understanding human mental representation is necessary for understanding human perception, cognition, decision making, and action. Mental representations play an important role in mediating adaptive behavior, and form the basis for the cognitive processes of generalization, inference and learning. Different assumptions regarding the nature and form of mental representation lead to different constraints on formal models of these processes. For this reason, Pinker (1998) argues that "pinning down mental representation is the route to rigor in psychology" (p. 85). Certainly, it is important that cognitive models use principled mental representations, since the *ad hoc* definition of stimuli on the basis of intuitive reasonableness is a highly questionable practice (Brooks 1991, Komatsu 1992, Lee 1998).

One appealing and widely used approach for deriving stimulus representations is to base them on measures of stimulus similarity. Following Shepard (1987), similarity may be understood as a measure of the degree to which the consequences of one stimulus generalize to another, and so it makes adaptive sense to give more similar stimuli mental representations that are themselves more similar. For a domain with n stimuli, similarity data take the form of an $n \times n$ similarity matrix, $\mathbf{S} = s_{ij}$, where s_{ij} is the similarity of the ith and jth stimuli. The goal of similarity-based representation is then to define stimulus representations that, under a given similarity model, capture the constraints implicit in the similarity matrix by approximating the data.

Goldstone's (in press) recent review identifies four broad model classes for stimulus similarity: geometric, featural, alignment-based, and transformational. Of these, the two most widely used approaches are the geometric, where stimuli are represented in terms of their values on different dimensions, and the featural, where stimuli are represented in terms of the presence or absence of weighted features. The geometric approach is most often used in formal models of cognitive processes, partly because of the ready availability of techniques such as multidimensional scaling (e.g., Kruskal 1964; see Cox & Cox 1994 for an overview), which generate geometric representations from similarity data. The featural approach to stimulus representation, however, is at least as important as the geometric approach, and warrants the development of techniques analogous to multidimensional scaling.

Accordingly, this paper describes an algorithm that generates featural representations from similarity data. The optimization processes used in the algorithm are standard ones, and could almost certainly be improved. In this regard, we draw on Shepard and Arabie's (1979) distinction between the psychological model that is being fit, and the algorithm that does the fitting. We make no claims regarding the significance of the algorithm itself (and certainly do not claim it is a model of the way humans learn mental representations), but believe that the psychological representational model that it fits has three important properties. First, it allows for the arbitrary definition of features, avoiding the limitations of partitioning or hierarchical clustering. Second, it uses a more general model of featural stimulus similarity than has previously been considered. Third, it generates featural representations in a way that balances the competing demands of data-fit and representational complexity.

Featural Representation

Within a featural representation, stimuli are defined by the presence or absence of a set of saliency weighted features or properties. Formally, if a stimulus domain contains n stimuli and m features, a featural representation is given by the $n \times m$ matrix $\mathbf{F} = f_{ik}$, where

$$f_{ik} = \begin{array}{ll} 1 & \text{if stimulus } i \text{ has feature } k \\ 0 & \text{otherwise,} \end{array} \quad (1)$$

together with a vector $\mathbf{w} = (w_1, \ldots, w_m)$ giving the (positive) weights of each of the features.

The Contrast and Ratio Models

Tversky's (1977) Contrast Model and Ratio Model of stimulus similarity provide a rich range of possibilities for generating featural representations that have been significantly under-utilized. Using the assumption that the similarity between two stimuli is a function of their common and distinctive features, the Contrast Model measures stimulus similarity as:

$$\hat{s}_{ij} = \theta F(\mathbf{f}_i \quad \mathbf{f}_j) - \alpha F(\mathbf{f}_i - \mathbf{f}_j) - \beta F(\mathbf{f}_j - \mathbf{f}_i), \qquad (2)$$

where $\mathbf{f}_i \quad \mathbf{f}_j$ denotes the features common to the ith and jth stimuli, $\mathbf{f}_i - \mathbf{f}_j$ denotes the features present in the ith, but not the jth, stimulus, and $F(\,)$ is some monotonically increasing function. By manipulating the positive weighting hyper-parameters θ, α and β, different degrees of importance may given to the common and distinctive components. In particular, Tversky (1977) emphasizes the two extreme alternatives obtained by setting $\theta = 1, \alpha = \beta = 0$ (common features only), and $\theta = 0, \alpha = \beta = 1$ (distinctive features only). A different approach is given by the Ratio Model, where similarity takes the form:

$$\hat{s}_{ij} = \frac{\theta F(\mathbf{f}_i \quad \mathbf{f}_j)}{\theta F(\mathbf{f}_i \quad \mathbf{f}_j) \quad \alpha F(\mathbf{f}_i - \mathbf{f}_j) \quad \beta F(\mathbf{f}_j - \mathbf{f}_i)}. \qquad (3)$$

While the Contrast Model and the Ratio Model provide great flexibility for measuring similarity across featural representations, the only established techniques for generating the representations from similarity data are additive clustering algorithms (e.g., Arabie & Carroll 1980; Lee 1999, in press; Mirkin 1987; Shepard & Arabie 1979; Tenenbaum 1996), which rely exclusively on the common features version of the Contrast Model. This means that only one special case of one of these approaches has been used as the basis of a practical technique for generating representations.

The paucity of available techniques is serious, given the recognition (e.g., Goodman 1972; Rips 1989; see Goldstone 1994 for an overview) that similarity is not a unitary phenomenon, and the way in which it is measured may change according to different cognitive demands. Direct empirical evidence that featural similarity judgments can place varying emphasis on common and distinctive features is provided by the finding that items presented in written form elicit common feature-weighted judgments, whereas pictures tend to be rated more in terms of distinctive features (Gati & Tversky 1984; Tversky & Gati 1978).

A Symmetric Contrast Model

Although the Contrast Model has three hyper-parameters, α and β remain distinct only when $s_{ij} = s_{ji}$. While it is certainly the case that real world domains display asymmetric similarity, modeling techniques based on similarity data generally assume that similarity is symmetric. Further, if the similarity ratings are assumed to lie between 0 and 1, the remaining hyper-parameters α and θ can be incorporated into one parameter, $\rho = \theta \quad (\theta \quad \alpha)$, which represents the relative weighting of common and distinctive features, with $0 \quad \rho \quad 1$. Setting the functional form $F(\,)$ using the same 'sum of saliency weights' approach as additive clustering yields the similarity model

$$\hat{s}_{ij} \quad = \quad \rho \sum_k w_k f_{ik} f_{jk} - \frac{1-\rho}{2} \quad \sum_k w_k f_{ik} \left(1 - f_{jk}\right)$$

$$\sum_k w_k (1 - f_{ik}) f_{jk} \qquad c. \qquad (4)$$

It is this symmetric version of the Contrast Model that is used in this paper to develop general featural representations. It allows for any relative degree of emphasis to be placed on common and distinctive features and, in particular, subsumes the additive clustering model ($\rho = 1$) and the distance-based feature-matching similarity model ($\rho = 0$). Technically, it is worth noting that the additive constant c used in additive clustering, which is added to all pairwise similarity estimates in both additive clustering and Contrast Model clustering representations, is not treated as a cluster, and thus is not weighted by ρ.

Limiting Representational Complexity

Shepard and Arabie (1979) have noted that the ability to specify large numbers of features and set their weights allows any similarity matrix to be modeled perfectly by a featural representation using the common features version of the Contrast Model. The same is true for the majority of Tversky's (1977) similarity models, and is certainly true for Eq. (4). While the representational power to model data is desirable, the introduction of unconstrained feature structures with free parameters detracts from fundamental modeling goals, such as the achievement of interpretability, explanatory insight, and the ability to generalize accurately beyond given information (Lee 2001a).

This means that techniques for generating featural representations from similarity data must balance the competing demands of maximizing accuracy and minimizing complexity, following the basic principle of model selection known as 'Ockham's Razor' (Myung & Pitt 1997). Data precision must also be considered, since precise data warrants a representation being made more detailed to improve data-fit, while noisy data does not.

In practice, this means that featural representations should not be derived solely on the basis of how well they fit the data, as quantified by a measure such as the variance accounted for,

$$\text{VAF} = 1 - \frac{\sum_{i<j}(s_{ij} - \hat{s}_{ij})^2}{\sum_{i<j}(s_{ij} - \bar{s})^2}, \qquad (5)$$

where $\bar{s}$ is the arithmetic mean of the similarity data. Rather, some form of complexity control must be used

to balance data-fit with model complexity. Most established algorithms strike this balance in unsatisfactory ways, either pre-determining a fixed number of clusters (e.g., Shepard & Arabie 1979; Tenenbaum 1996), or predetermining a fixed level of representational accuracy (e.g., Lee 1999).

Recently, Lee (in press) has applied the Bayesian Information Criterion (BIC: Schwarz 1978) to limit the complexity of additive clustering representations. Unfortunately, an important limitation of the BIC is that it equates model complexity with the number of parameters in the model. While this is often a reasonable approximation, it neglects what Myung and Pitt (1997) term the 'functional form' component of model complexity. For featural representations, parametric complexity is simply the number of features used in a representation. Functional form complexity, however, considers the feature structure **F**, and is sensitive to the patterns with which stimuli share features (see Lee 2001a), as well as any difference arising from the relative emphasis given to common and distinctive features.

It is important to account for functional form complexity with featural representational models that can vary their emphasis on common and distinctive features. Figure 1 shows the results of fitting featural representations, assuming different levels of ρ, on similarity data that were generated using either entirely common features ($\rho = 1$), entirely distinctive features ($\rho = 0$), or an even balance of the two ($\rho = 0.5$). These results are averaged across five different similarity matrices, each based on a five-feature representation, and show one standard error about the mean level of fit.

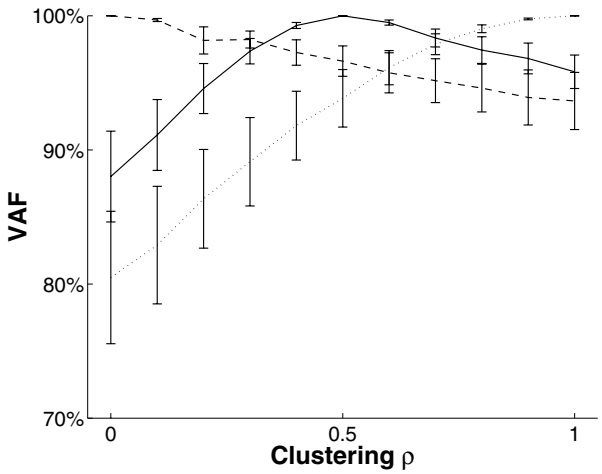


Figure 1: The change in VAF value, as a function of the assumed balance between common and distinctive features, for the entirely common (dotted line), entirely distinctive (dashed line) and balanced (solid line) similarity data.

As expected, the best-fitting featural representations have ρ values matching those that generated the data. More interestingly, Figure 1 shows that the level of fit for the entirely common features data deteriorates more rapidly than for the entirely distinctive features data when the wrong ρ value is assumed. Similarly, for the evenly balanced data, the fit is greater when too much emphasis is placed on common features in the assumed similarity model. These results imply that common features-weighted models are more able to fit data when they are wrong than are distinctive features-weighted models. In the language of model complexity, the common features functional form is more flexible than the distinctive features functional form, and this extra complexity improves the fit of incorrect models. For this reason, it is important to derive featural representations using a measure that is sensitive to functional form complexity.

A Geometric Complexity Criterion

Myung, Balasubramanian, and Pitt (2000) have recently developed a measure called the Geometric Complexity Criterion (GCC) that constitutes the state-of-the-art in accounting for both fit and complexity in model selection. The basic idea is to define complexity in terms of the number of distinguishable data distributions that the model can accommodate through parametric variation, with more complicated models being able to index more distributions than simple ones. Using Tenenbaum's (1996) probabilistic formulation of the data-fit of a featural model, and extending Lee's (2001a) derivation of the Fisher Information matrix for the common features case of the Contrast Model, it is a reasonably straightforward exercise to derive a GCC for the current similarity model. The final result is:

$$\text{GCC} = \frac{1}{2s^2}\sum_{i<j}(s_{ij}-\hat{s}_{ij})^2 \quad \frac{m}{2}\ln\frac{n(n-1)}{4\pi s^2} \quad \frac{1}{2}\ln\det G, \tag{6}$$

where s denotes an estimate of the inherent precision of the data (see Lee 2001b), m is the number of features, n is the number of stimuli, and G denotes the $m \times m$ *complexity matrix* for the feature structure. The xy-th cell of the complexity matrix is given by,

$$\sum_{i<j} e_{ijx}e_{ijy} \tag{7}$$

where e_{ijx} equals ρ if x is a common feature, $-(1-\rho)$ 2 if x is a distinctive feature, and 0 if neither i nor j possesses the feature x.

An interesting aspect of the complexity matrix, and the GCC measure as a whole, is that it is independent of the parameterization of the model. That is, the complexity of a featural representation is dependent only on the feature structure, and not the saliencies assigned to the features. We should make two technical points about the GCC. First, this derivation is based on the assumption that ρ is a fixed property of a model, and not a free parameter.

An alternative would be to modify the GCC so that it accommodated ρ as a model parameter. Second, since the additive constant is not weighted by ρ, the terms in the complexity matrix corresponding to the additive constant behave as if $\rho = 1$.

Algorithm

In developing an algorithm to fit featural representations using the Contrast Model, we were guided by the successful additive clustering algorithm reported by Lee (submitted). Basically, the algorithm works by 'growing' a featural representation, starting with a one-feature model, and continually adding features while this leads to improvements in the GCC measure. For any fixed number of features, the search for an appropriate assignment of stimuli to features is done using stochastic hill-climbing, with the best-fitting weights being determined using a standard non-negative least squares algorithm (Lawson & Hanson 1974). The algorithm terminates once the process of adding features leads to representations with GCC values that are more than a prespecified constant above the best previously found, and the featural representation with the minimum GCC value is returned.

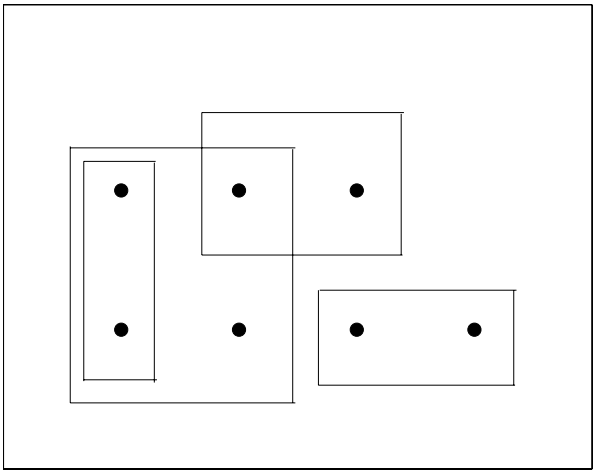


Figure 2: The artificial featural representation containing seven stimuli and four features.

To test the ability of this optimization algorithm to fit similarity data, we examined its ability to recover a known featural representation. This representation had seven stimuli and four features, and included partitioning, nested, and overlapping clusters, as shown in Figure 2. Using this representation, similarity data were generated assuming entirely common features, entirely distinctive features, or an even balance between the two. Feature weights were chosen at random subject to the constraint that they resulted in positive similarity values. Each of the similarity values was perturbed by adding noise that was independently drawn from a Normal distribution with mean 0 and standard deviation 0.05.

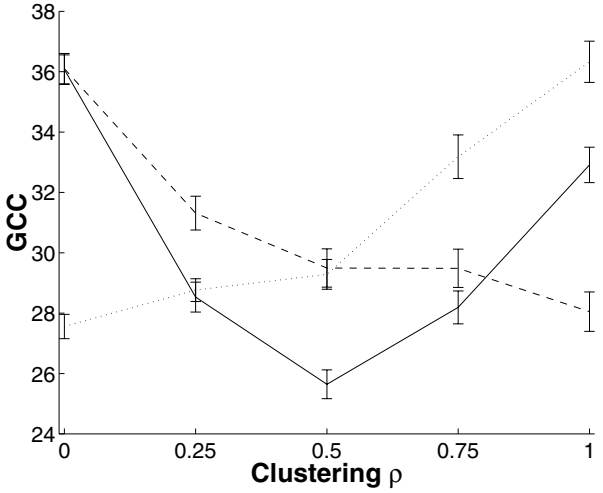


Figure 3: The change in GCC value, as a function of the assumed balance between common and distinctive features, for the entirely common (dotted line), entirely distinctive (dashed line) and balanced (solid line) similarity data.

The algorithm was applied to this similarity data under different assumptions regarding the balance between common and distinctive features, using ρ values of 0, 0.25, 0.5, 0.75 and 1. In calculating the GCC measure, a data precision value of $\sigma = 0.05$ was assumed, in accordance with the known level of noise. Figure 3 summarizes the results of 10 runs of the algorithm for each of the three similarity conditions, across all of the assumed ρ values. The mean GCC value of the 10 derived representations is shown, together with error bars showing one standard error in both directions.

Figure 3 shows that the GCC is minimized at the correct ρ value for all three similarity conditions. An examination of the derived representation revealed that the correct featural representation was recovered 25 times out of 30 attempts: nine times out of ten for the entirely distinctive data, and eight times out of ten for the evenly balanced and the entirely common data. It is interesting to note that Figure 3 is far more symmetric than Figure 1, suggesting that the GCC has successfully accounted for the differences in functional form complexity between the common and distinctive feature approaches to measuring similarity.

Additional Monte Carlo simulations with other featural representations, based on particular structures reported by Tenenbaum (1996, Table 1) and Lee (1999, Table 5), also suggested that the algorithm is capable of recovering known configurations when more stimuli or more features are involved, although problems with local minima are encountered more frequently.

Table 1: Representation of Rosenberg and Kim's (1975) kinship terms domain.

STIMULI IN CLUSTER	WEIGHT
aunt uncle niece nephew cousin	0.319
granddaughter grandson grandmother grandfather	0.291
mother daughter grandmother granddaughter aunt niece sister	0.222
sister brother cousin	0.221
father son grandfather grandson uncle nephew brother	0.208
mother father daughter son sister brother	0.163
mother father daughter son	0.136
daughter son granddaughter grandson niece nephew sister brother	0.128
mother father grandmother grandfather aunt uncle sister brother	0.091
additive constant	0.563
VARIANCE ACCOUNTED FOR	92.7%

An Illustrative Example

To demonstrate the practical application of the algorithm, we used the averaged similarity data reported by Rosenberg and Kim (1975), which measures similarity of English kinship terms. A data precision estimate of $s = 0.09$ was made based on the sample standard deviation of the individual matrices. Since the data was obtained by having participants sort items into different stacks, we might expect a model that provides a weighting of common and distinctive features to provide a better fit than one allowing only for common features. Using ρ values of 0, 0.1, 0.2, ..., 1.0, the representation with the minimum GCC was found at $\rho = 0.4$.

This representation contained the nine features detailed in Table 1, and explained 92.7% of the variance in the data. Interpreting most of the features in Table 1 is straightforward, since they essentially capture concepts such as 'male', 'female', 'nuclear family', 'extended family', 'grandparents', 'descendants', and 'progenitors'. While this representation is very similar to the nine-feature representation generated by additive clustering (Lee submitted, Figure 2), it explains more of the variance in the data, suggesting that participants did indeed use both common and distinctive features in assessing similarity.

Conclusion

We have developed, tested, and demonstrated an algorithm that generates featural stimulus representations from similarity data. Unlike previous additive clustering approaches, the algorithm uses a symmetric version of Tversky's (1977) Contrast Model that measures similarity in terms of both common and distinctive features. A particular strength of the algorithm is its use of the Geometric Complexity Criterion to guide the generation process, which allows the desire for data-fit to be balanced with the need to control representational complexity. Importantly, this criterion is sensitive to the functional form complexity of the similarity model, preventing an overemphasis on the inherently more complicated common features approach.

In terms of future work, it should be acknowledged that the symmetric version of the Contrast Model is certainly not the only possibility for combining common and distinctive features approaches to measuring similarity. Tenenbaum and Griffiths (in press) provide a compelling argument for the use of the Ratio Model in the context of their Bayesian theory of generalization. It would also be worthwhile to examine featural representations where each feature is assumed to operate using entirely an distinctive or an entirely common approach. The distinctive similarity features would be those that globally partition the entire stimulus set, as for the feature 'male', which implies the existence of the complementary feature 'female'. The (more prevalent) common similarity features would be those that captured shared properties, such as eye or hair color, where no broader implications are warranted.

Acknowledgments

This article was supported by a Defence Science and Technology Organisation scholarship awarded to the first author. We wish to thank several referees for helpful comments on an earlier version of this paper.

References

Arabie, P., & Carroll, J.D. (1980). MAPCLUS: A mathematical programming approach to fitting the ADCLUS model. *Psychometrika 45*(2), 211–235.

Brooks, R.A. (1991). Intelligence without representation. *Artificial Intelligence 47*, 139–159.

Cox, T.F., & Cox, M.A.A. (1994). *Multidimensional scaling*. London: Chapman and Hall.

Gati, I., & Tversky, A. (1984). Weighting common and distinctive features in perceptual and conceptual judgments. *Cognitive Psychology 16*, 341–370.

Goldstone, R.L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition 52*, 125–157.

Goldstone, R.L. (in press). Similarity. In R.A. Wilson & F.C. Keil (Eds.) *MIT Encyclopedia of the Cognitive Sciences*. Cambridge, MA: MIT Press.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and Projects*, pp. 437–446. New York: Bobbs-Merrill.

Lawson, C.L., & Hanson, R.J. (1974).*Solving Least Squares Problems*. Englewood Cliffs, NJ: Prentice-Hall.

Lee, M.D. (1998). Neural feature abstraction from judgments of similarity. *Neural Computation, 10* (7), 1815-1830.

Lee, M.D. (1999). An extraction and regularization approach to additive clustering. *Journal of Classification, 16* (2), 255-281.

Lee, M.D. (2001a). On the complexity of additive clustering models. *Journal of Mathematical Psychology, 45* (1), 131-148.

Lee, M.D. (2001b). Determining the dimensionality of multidimensional scaling models for cognitive modeling. *Journal of Mathematical Psychology, 45* (1), 149-166.

Lee, M.D. (in press). A simple method for generating additive clustering models with limited complexity. *Machine Learning*.

Lee, M.D. (submitted). *Generating additive clustering models with limited stochastic complexity*. Manuscript submitted for publication.

Komatsu, L.K. (1992). Recent views of conceptual structure. *Psychological Bulletin 112*(3), 500–526.

Kruskal, J.B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29* (1), 1–27.

Mirkin, B.G. (1987). Additive clustering and qualitative factor analysis methods for similarity matrices. *Journal of Classification 4*, 7–31.

Myung, I.J., Balasubramanian, V., & Pitt, M.A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences USA 97*, 11170–11175.

Myung, I.J., & Pitt, M.A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin and Review 4*(1), 79–95.

Pinker, S. (1998). *How the mind works*. Great Britain: The Softback Preview.

Rips, L.J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning*, pp. 21–59. New York: Cambridge University Press.

Rosenberg, S., & Kim, M.P. (1975). The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research 10*, 489–502.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Shepard, R.N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317–1323.

Shepard, R.N., & Arabie, P. (1979). Additive clustering representations of similarities as combinations of discrete overlapping properties. *Psychological Review 86*(2), 87–123.

Tenenbaum, J.B. (1996). Learning the structure of similarity. In D.S. Touretzky, M.C. Mozer, & M.E. Hasselmo (Eds.), *Advances in Neural Information Processing Systems*, Volume 8, pp. 3–9. Cambridge, MA: MIT Press.

Tenenbaum, J.B., & Griffiths, T.L. (in press). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences 24*(2).

Tversky, A. (1977). Features of similarity. *Psychological Review 84*(4), 327–352.

Tversky, A., & Gati, I. (1978). Studies of similarity. In E. Rosch and B.B. Lloyd (Eds.), *Cognition and Categorization*, pp. 79–98. Hillsdale, NJ: Wiley.

Active inference in concept learning

Jonathan D. Nelson (jnelson@cogsci.ucsd.edu)*
Joshua B. Tenenbaum (jbt@psych.stanford.edu)^
Javier R. Movellan (movellan@cogsci.ucsd.edu)*

*Cognitive Science Department, UCSD
La Jolla, CA 92093-0515

^Psychology Department, Stanford University
Stanford, CA 94305

Abstract

People are active experimenters, constantly seeking new information relevant to their goals. A reasonable approach to active information gathering is to ask questions and conduct experiments that minimize the expected state of uncertainty, or maximize the expected information gain, given current beliefs (Fedorov, 1972; MacKay, 1992; Oaksford & Chater, 1994). In this paper we present results on an exploratory experiment designed to study people's active information gathering behavior on a concept learning task. The results of the experiment suggest subjects' behavior may be explained well from the point of view of Bayesian information maximization.

Introduction

In scientific inquiry and in everyday life, people seek out information relevant to perceptual and cognitive tasks. Whether performing experiments to uncover causal relationships, saccading to informative areas of visual scenes, or turning towards a surprising sound, people actively seek out information relative to their goals.

Consider a person learning a foreign language, who notices a particular word, "tikos," used to refer to a baby moose, a baby penguin, and a baby cheetah. Based on those examples, she may attempt to discover what tikos really means. Logically, there are an infinite number of possibilities. For instance, tikos could mean *baby animals*, or simply *animals*, or even *baby animals and antique telephones*. Yet a few examples are often enough for human learners to form strong intuitions about what meanings are most likely.

Suppose the learner could point to a baby duck, an adult duck, or an antique telephone, to inquire whether that object is "tikos." What question would she ask? Why do we think that pointing to the telephone is not a good idea, even though from a logical point of view, a phone could very well be tikos? In this paper we present a normative theoretical framework, to try to predict the questions people ask in concept learning tasks (Fedorov, 1972; MacKay, 1992; Oaksford & Chater, 1994).

A Bayesian concept learning model

In the approach presented here, we evaluate questions in terms of their information value. Formally, information is defined with respect to a probability model. Here we use a Bayesian framework in the sense that we model internal beliefs as probability distributions. In order to quantify the information value (in bits) of a person's questions, we first need a model of her beliefs, and the way those beliefs are updated as new information is obtained. Tenenbaum (1999, 2000) provides such a model of people's beliefs, for a number concept learning task. While Tenenbaum (1999, 2000); and the first and last authors of the present paper, in a pilot study, found that his model described subjects' beliefs well, there were some deviations between model predictions and subjects' beliefs. The concept learning model used in the present study, which we describe below, is based on Tenenbaum's original model, but extended in ways that reduce previously observed deviations between model predictions and study participants' beliefs.

We formalize the concept learning situation described by the number concept model using standard probabilistic notation: random variables are represented with capital letters, and specific values taken by those variables are represented with small letters. The random variable C represents the correct hidden concept on a given trial. This concept is not directly observable by study participants; rather, they infer it on the basis of example numbers that are consistent with the true concept. Notation of the form *"C=c"* is shorthand for the event that the random variable C takes the specific value c, e.g. that the correct concept (or "hypothesis") is *prime numbers*. We represent the examples given to the subjects by the random vector X. The subject's beliefs about which concepts are probable prior to the presentation of any examples is represented by the probability function $P(C=c)$. The subject's updated belief about a concept's probability, after she sees the

examples X=x, is represented by *P(C=c|X=x)*. For example, if *c* is the concept *even numbers* and *x* the numbers "2, 6, 4", then *P(C=c|X=x)* represents the subject's posterior probability that the correct concept is *even numbers*, given that 2, 6, and 4 are positive examples of that concept. Study participants are not explicitly given the true hidden concept; rather, they infer it from examples of numbers that are consistent with the true concept.

The number concept model includes both *arithmetic* and *interval* concepts. Interval concepts are sets of consecutive integers between *n* and *m*, where $1 \le n \le 100$, and $n \le m \le 100$, such as *numbers between 5 and 8*, and *numbers between 10 and 35*. Thus, there are 5050 interval concepts. Arithmetic concepts include *odd numbers*, *even numbers*, *square numbers*, *cube numbers*, *prime numbers*, *multiples of n* ($3 \le n \le 12$), *powers of n* ($2 \le n \le 10$), and *numbers ending in n* ($1 \le n \le 9$). There are 33 arithmetic concepts.

Inferences are made with respect to the following model of how examples are generated: A concept is first chosen at random according to a prior probability distribution. The prior probability distribution of the model is designed to reflect the human intuition that a concept like *multiples of 10* is more plausible than a concept like *multiples of 10 except 30*. A portion of total prior probability is divided evenly into the arithmetic concepts, with the exception of *even numbers* and *odd numbers*. To reflect the higher salience of the concepts *even numbers* and *odd numbers*, each of those concepts is given five times the prior probability of the other arithmetic concepts. Among the interval concepts, prior probability is apportioned according to the Erlang distribution

$$P(H = h) \propto \frac{|h|}{\sigma^2} e^{\frac{-|h|}{\sigma}}$$

according to the concept's size $|h|$. (The concept *numbers between 15 and 30* is size 16.) Sigma gives the optimal interval length. In the simulations described in this paper we set σ to 15, although in principle, σ is a free parameter to fit to the data. Interval concepts of a given length, such as *numbers between 25 and 35*, and *numbers between 89 and 99*, receive the same prior probability, irrespective of their endpoints.

Once a concept is chosen, examples are randomly and independently generated, with equal probability, from the set of numbers in that concept. Thus, the likelihood of a particular vector of m examples X=x, given the concept h,

$$P(X = x \mid H = h) = \frac{1}{|h|^m},$$

if all m examples are in the concept *h*, and zero otherwise.

This generating assumption reflects the human intuition that although a given set of example numbers is typically compatible with more than one concept, it may be more representative of some concepts than others. For instance, although the example numbers 60, 80, 10, and 30 are compatible with both *multiples of 10* and *multiples of 5*, that set of numbers is a better example of the concept *multiples of 10* than it is of the concept *multiples of 5*, because it is much more likely to be observed as a random sample from the more specific hypothesis *multiples of 10*.

The generative model described above can be used to compute the probability that a new element *y* belongs to the hidden concept *C* given the examples in *x*:

$$P(y \in C \mid X = x) = \frac{\sum_{h: y \in h} P(X = x \mid H = h) P(H = h)}{\sum_{h} P(X = x \mid H = h) P(H = h)}$$

An ideal concept learning model would assign some prior probability to every possible concept, according to each concept's plausibility to human learners. The main difference between the concept learning model used in the current paper, and the model introduced in Tenenbaum (1999, 2000), is our inclusion of a large number of random "exception" concepts, which are formed by replicating and slightly changing, or "mutating," concepts from the basic model. Here, we include 50,830 exception hypotheses -- on average, 10 exception concepts for each concept in the basic model. To form an exception concept (or "hypothesis"), a concept is first picked from the basic model, according to the prior probability of concepts in the basic model. We include a parameter μ for the average number of changes to the original concept, and divide these changes equally, on average, into additions of new numbers and exclusions of existing numbers. The probability of each existing number being excluded from a concept is $\frac{\mu}{2|h|}$, and the probability of each currently excluded number (between 1 and 100) being added to the concept is $\frac{\mu}{2(100 - |h|)}$.

Each exception hypothesis receives a constant share of the total proportion of prior probability assigned to the exception hypotheses. In the simulation of the model reported in this paper, 60% of prior probability

was assigned to the exception hypotheses, and μ was set to 6. It takes approximately 30 minutes to simulate the set of trials in the study, for any setting of model parameters, and we are just beginning to explore the parameter space. Early exploration suggests that a wide range of parameters in the extended number concept model can improve on the basic model's correspondence to human beliefs.

Information-maximizing sampling

In the experiment reported in this paper, we allowed subjects to actively ask questions about number concepts, instead of making inferences solely on the basis of the examples given to them. For example, on one trial the subject was given the number 16 as an example of the hidden underlying concept, and then was allowed to test another number, to find out whether it was also consistent with the true, hidden concept.

In our formalism, the binary random variable Y_n represents whether the number n is a member of the correct concept. For example, $Y_8=1$ represents the event that 8 is an element of the correct, hidden concept, and $Y_8=0$ the event that 8 is not in that concept. Asking "is the number n an element of the concept?" is equivalent to finding the value taken by the random variable Y_n, in our formalism.

We evaluate how good a question is in terms of the information about the correct concept expected for that question, given the example vector $X=x$. The expected information gain for the question "Is the number n an element of the concept?" is calculated with respect to the learner's beliefs, as approximated with the extended number concept model described above. Formally, expected information gain is given by the following formula:

$$I(C, Y_n \mid X = x) = H(C \mid X = x) - H(C \mid Y_n, X = x),$$

where the uncertainty (entropy) about the hidden concept C given the example numbers in x,

$$H(C \mid X = x) = -\sum_c P(C = c \mid X = x) \log_2 P(C = c \mid X = x),$$

and the expected remaining uncertainty about the hidden concept C, given the example numbers in x and the answer to the question Y_n:

$$H(C \mid Y_n, X = x) = -\sum_{v=0}^{1} P(Y_n = v \mid X = x) \sum_c P(C = c \mid Y_n = v, X = x) \log_2 P(C = c \mid Y_n = v, X = x)$$

We consider only binary questions, of the form "is n consistent with the concept?" so the maximum information value of any question in our experiment is one bit. Note how information value of questions is relative to subjects' internal beliefs, which we approximate here by using the expanded number concept learning model. An information-maximizing strategy prescribes asking the question with the highest expected information gain, e.g., the question that minimizes the expected entropy, over all concepts.

Another strategy of interest is confirmatory sampling, which consists of asking questions whose answers are most likely to confirm current beliefs. In other domains it has been proposed that people have a bias to use confirmatory strategies, regardless of their information value (Klayman & Ha, 1987; Popper, 1959; Wason, 1960).

The active sampling concept game

Twenty-nine undergraduate students, recruited from Cognitive Science Department classes at the University of California, San Diego, participated in the experiment. Subjects gave informed consent, and received either partial course credit for required study participation, or extra course credit, for their participation. The experiment began with the following instructions:

> Often it is possible to have a good idea about the state of the world, without completely knowing it. People often learn from examples, and this study explores how people do so. In this experiment, you will be given examples of a hidden number rule. These examples will be randomly chosen from the numbers between 1 and 100 that follow the rule. The true rule will remain hidden, however. Then you will be able to test an additional number, to see if it follows that same hidden rule. Finally, you will be asked to give your best estimation of what the true hidden rule is, and the chances that you are right. For instance, if the true hidden rule were "*multiples of 11*," you might see the examples 22 and 66. If you thought the rule were "*multiples of 11*," but also possibly "*even numbers*," you could test a number of your choice, between 1-100, to see if it also follows the rule.

On each trial subjects first saw a set of examples from the correct concept. For instance, if the concept were even numbers, subjects might see the numbers "2, 6, 4" as examples. Subjects were then given the opportunity to test a number of their choice. Subjects were given feedback on whether the number they tested was an element of the correct concept.

We wrote a computer program to simulate the expanded number concept model, and to compute the information value of each possible question, given each set of examples. By considering beliefs and questions together, we may evaluate the information value of participants' questions, as well as that of information-maximizing and confirmatory sampling strategies. We define the confirmatory strategy as testing the number (excluding the examples) that has the highest posterior probability, as given by the extended number concept

model, of being consistent with the correct hidden concept.

Results

We discuss two types of trials, grouped according to the posterior beliefs of the extended number concept model, after all the example numbers have been seen. These results should be considered preliminary, as 29 data points on each trial are not sufficient for estimation of statistically reliable sampling distributions over the range of possible queries from 1 to 100.

Arithmetic trials

On some trials, the model is dominated by arithmetic concepts, and exception hypotheses based on arithmetic concepts. On each of these trials, good agreement between a number's information value and subjects' propensity to sample that number was observed. The information value of the confirmatory strategy was near to that of the information-maximizing strategy on these trials.

Consider the trial with the examples 81, 25, 4, and 36, in which the concept with the highest posterior probability is *square numbers*. Generalization behavior of the model, and beliefs of subjects, are shown in Figure 1. Note that the model and subjects alike assign certain, or near certain, probability to each of the example numbers, but less than certain probability to the other square numbers. Relative to the model's beliefs, the most informative numbers to test are non-example square numbers, such as 9, 16, 49, 64, or 100 (Figure 2). In fact, 20 of 29 subjects tested one of these numbers. Other subjects' samples do not show a clear pattern, except for testing the number 10 (5 of 29 subjects), which is unpredicted.

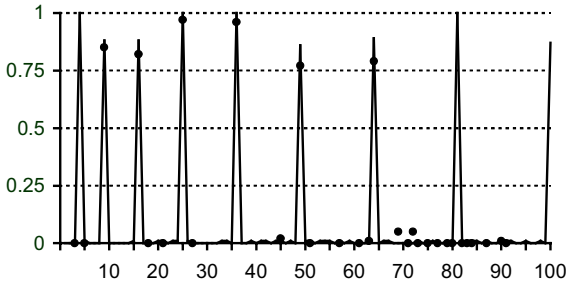


Figure 1. Generalization probabilities, given the examples 81, 25, 4, and 36. Model probabilities are given by the line. Subjects' probabilities, for the 30 probe numbers subjects rated, are given with circles.

Good agreement between subjects' samples and rated information value is also observed on the trial with the examples 16, 8, 2, and 64. The most informative numbers to test are non-example powers of two, 4 or 32. Most (16/29) subjects tested these numbers.

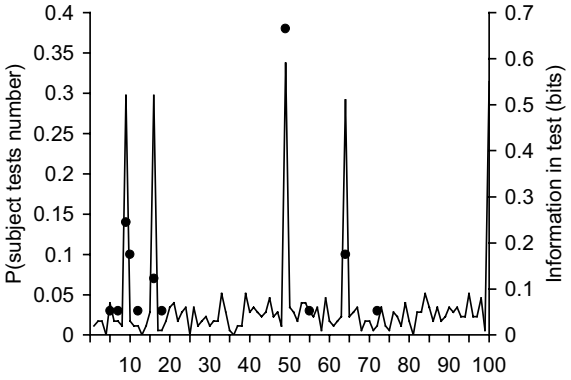


Figure 2. Information value of questions (line), and subjects' questions (circles), given the examples 81, 25, 4, and 36.

Finally, we may consider the trial with the examples 60, 80, 10, and 30, in which the hypothesis *multiples of 10* receives the highest posterior probability; multiples of 5 also receive moderate probability (Figure 4). On this trial, non-example multiples of 10, such as 20, and odd multiples of five, have the highest information value. Multiples of 10 were tested by 21 of 29 subjects; an additional 5 subjects tested odd multiples of five

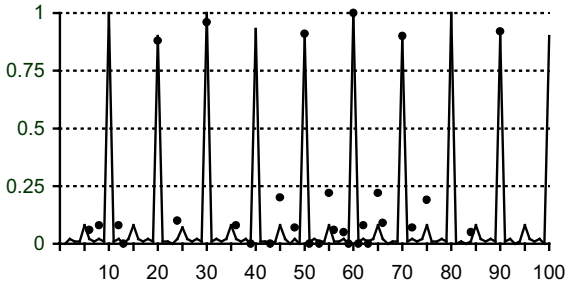


Figure 3. Generalization probabilities given the examples 60, 80, 10, and 30.

The difference between the first two arithmetic trials, and the trial with the examples 60, 80, 10, and 30 appears to be that a clear alternate hypothesis -- multiples of five -- receives moderate posterior probability in the multiples of 10 trial, but not on the other trials.

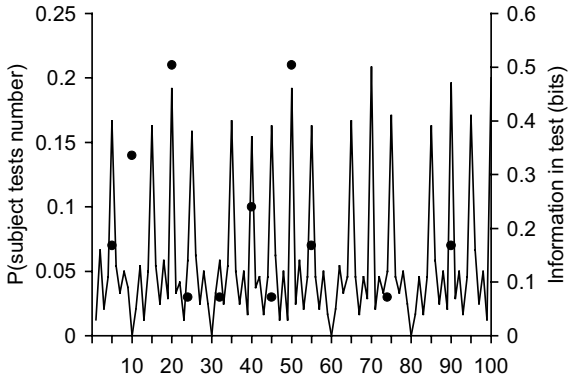


Figure 4. Information value of questions, and subjects' questions, given the examples 60, 80, 10, and 30.

Interval trials

On these trials, several examples of numbers of similar magnitude, such as 16, 23, 19, and 20, are given (these numbers are points where model probabilities are 1.00, Figure 5, and Figure 7). The model is certain that the example numbers themselves are consistent with the true concept. The model is fairly sure that non-example numbers within the range spanned by the examples, like 17, 18, 21, and 22, are consistent with the true concept. Finally, the model assigns decreasing probability to numbers as they move away from the range of observed examples (Figure 5).

It should be noted that there is some variability from one run of the model to the next. The general pattern of results, however, holds from run to run. In particular, (1) numbers slightly outside of the range of the observed examples are most informative, (2) information value of numbers decreases with increasing distance from the observed examples, and (3) there is moderate information value in non-example numbers within the range of observed examples.

Most subjects tested numbers outside of, but near the observed examples (Figure 6). About one-third of subjects tested (non-example) numbers within the range spanned by the examples. On the other interval trials -- with example numbers 60, 51, 57, and 55; and 81, 98, 96, 93 (illustrated in Figure 7 and Figure 8)-- similar patterns emerged.

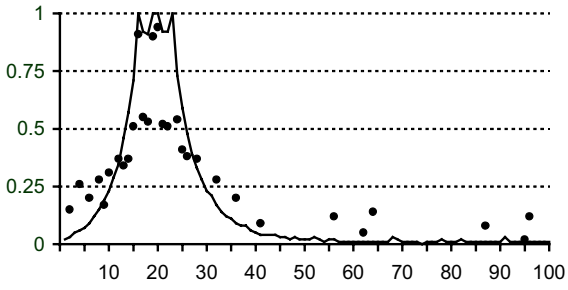


Figure 5. Generalization probabilities, given the examples 16, 23, 19, and 20.

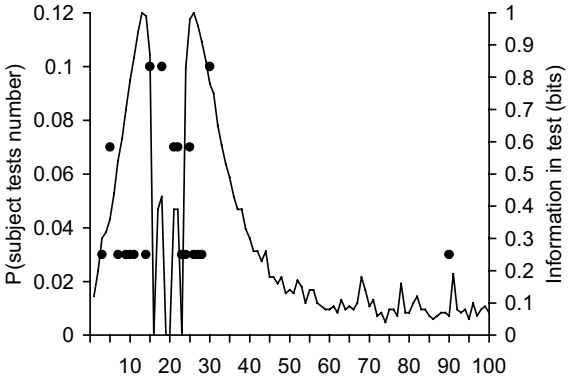


Figure 6. Information value of questions, and subjects' questions, given the examples 16, 23, 19, and 20.

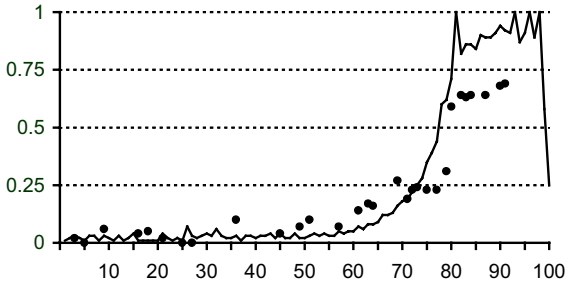


Figure 7. Generalization probabilities, given the examples 81, 98, 96, and 93.

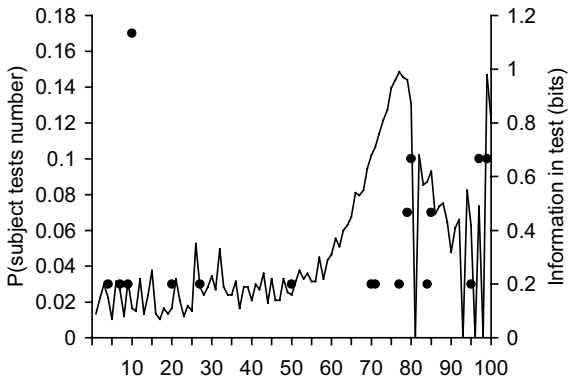


Figure 8. Information value of questions, and subjects' questions, given the examples 81, 98, 96, and 93.

Discussion

This paper presents work in progress to analyze active inference in concept learning from the point of view of the rational, probabilistic approach to cognition (Anderson, 1990). In the rational study of information-gathering behavior, the current research adds to existing analyses of Wason's (1966, 1968) selection task (Oaksford & Chater, 1994, 1998), and Wason's (1960) 2-4-6 task (Ginzburg & Sejnowski, 1996).

We found that a normatively inspired criterion of optimal sampling -- maximizing average information gain -- predicts human behavior well on a relatively unconstrained task. This result is strengthened by the fact that the extended number concept model we employed, as a proxy for subjects beliefs, was not originally developed with the goal of serving as a model for sampling. Nor were our extensions to it ad hoc. To the contrary, our extended model now has a better fit to data from earlier studies.

If rational theories of cognition are to explain thought and behavior in natural environments, then optimal sampling agents should also exhibit the systematic "biases" traditionally associated with human behavior. Indeed, we found that on many trials, a confirmatory sampling strategy approximates the information-maximizing strategy.

A final point is that whereas information gain, calculated with respect to the extended number concept model, predicts study participants' questions fairly well, information gain with respect to the original number concept model does not do so. This illustrates that particular queries are not informative or uninformative on their own, but only in relation to a particular probability model. To understand people's questions, or build artificial sampling systems that come closer to meeting human competence, developing appropriate probability models is critical.

Acknowledgments

Thanks to Gedeon Deák, Jeff Elman, Iris Ginzburg, Craig McKenzie, Terry Sejnowski, and three anonymous reviewers for their ideas; and Kent Wu, Dan Bauer and Jonathan Weh for their help in this research. J. Nelson was partially supported by a Pew graduate fellowship during this research.

References

Anderson, J. R. (1990). *The adaptive character of thought*. New Jersey: Erlbaum.

Fedorov, V. V. (1972). *Theory of optimal experiments*. New York: Academic Press.

Ginzburg, I.; Sejnowski, T. J. (1996). Dynamics of rule induction by making queries: transition between strategies. *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, 121-125.

Klayman, J.; Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94, 211-228.

MacKay, D. J. C. (1992). Information-based objective functions for active data selection. *Neural Computation*, 4, 590-604.

Oaksford, M.; Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101, 608-631.

Oaksford, M.; Chater, N. (1998). *Rationality in an uncertain world: Essays on the cognitive science of human reasoning*. UK: Erlbaum

Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.

Tenenbaum, J. B. (1999). *A Bayesian Framework for Concept Learning.* Ph.D. Thesis, MIT

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. In *Advances in Neural Information Processing Systems*, 12, Solla, S. A., Leen, T. K., Mueller, K.-R. (eds.), 59-65.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*. 12, 129-140.

Wason, PC (1966). Reasoning. In Foss, B (ed.), *New Horizons in Psychology*, pp. 135-151.

Wason, PC (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20, 273-281.

Addition as Interactive Problem Solving

Hansjörg Neth (NethH@Cardiff.ac.uk)
School of Psychology, Cardiff University
Cardiff CF10 3YG, Wales, United Kingdom

Stephen J. Payne (PayneS@Cardiff.ac.uk)
School of Psychology, Cardiff University
Cardiff CF10 3YG, Wales, United Kingdom

Abstract

Successful problem solving depends on a dynamic interplay of resources between agent, task, and task environment. To illuminate these interactions we studied how participants added a series of single-digit numbers presented on a computer screen. We distinguished between four different user interfaces, each implementing a different mode of interaction with the displayed addends: look only, point, mark, and move. By collecting and analysing complete interaction protocols we were able to integrate overall performance measures with fine-grained behavioural process data on the strategies engendered by the different user interfaces. We discovered reliable differences in the chosen sequences of addends, which can be understood in terms of the cost-benefit structures provided by the interactive resources of the user interfaces.

Introduction

Successful problem solving is an embedded and embodied process and crucially depends on a dynamic interplay of resources and constraints between agent, task, and task environment.

The importance of feedback loops, wherein actions on the world provide new information to the problem solver, was recognized in the earliest cognitive accounts of human problem solving (e.g. Miller, Galanter and Pribram, 1960; Newell and Simon, 1972). Yet until relatively recently, the interactive properties of the task environment have seldom been the focus of attention. Thus, the traditional literature on problem solving has been concerned primarily with planning, search strategies and heuristics (see e.g. Mayer, 1992, for an overview).

Recently, however, it has become increasingly clear to many investigators that interactions between mental processes and external objects play a crucial role in human problem solving. This interactive perspective has led to recent analyses of, for example: the importance of constraints provided by the properties of external representations (Larkin and Simon, 1987; Zhang and Norman, 1994; Zhang, 1997); the role of the display as a resource in human-computer interaction (Payne, 1991; Monk, 1998; Gray and Fu, 2001); the effect of the cost of implementing operators on the interplay between planning and action (O'Hara and Payne, 1998).

In this article, we extend this general approach to investigate the way in which the nature of available interactions in the task environment determines the discovery and use of strategies in a rather simple problem solving task: adding a series of numbers.

This work builds on the empirical work of Kirsh and colleagues (1995a, 1995b, Kirsh & Maglio, 1994), who, in a series of empirical studies, have shown that problem solvers often spontaneously manipulate the external world in order to reduce cognitive load.

In studying the interactive video game "Tetris", Kirsh and Maglio (1994) showed that expert players physically rotated falling pieces more than was required by their goal orientation. Kirsh (1995) demonstrated that people were reliably faster and more accurate at counting coins when they were allowed to move the coins around as they counted. Similarly, Maglio et al. (1999) found that people generated more anagrams when they were allowed to rearrange Scrabble tiles as they worked.

The experiment reported in the current article exemplifies an empirical approach that has three characteristics:

First, we study a problem solving task in which the atomic components are relatively simple and well understood, so that strategy differences (as well as outcome differences) may be easier to observe and explain.

Second, we design several user interfaces to the same problem, allowing subtle manipulations to the interactive resources that are available to problem solvers. This enables a more refined investigation of the relationship between resources and strategies.

Third, we independently manipulate problem complexity, so that we can assess relations between problem characteristics and interactive resources.

Interaction in Addition

Consider a simple serial addition like 1+2+9+7 presented, as here, linearly on a visual display. As with many cognitive tasks, we could solve this entirely "in our heads". But this does not warrant the conclusion that environmental interactions cannot play important roles.

Despite the linear presentation format, the law of commutativity allows us to add the four addends in any of 4!=24 different orders. Whilst all potential solution paths result in a total sum of 19 their cognitive demands may vary considerably. Within the context of this study, two sequences are of particular interest: By first adding 1 plus 9 before adding 2 and 7 to the result, one could exploit the fact that within the arabic base 10 number system the two addends 1 and 9 form what we call a *pair*,

i.e., they add up to the next bigger unit, a "round number". Likewise, someone might first add 1+2 but then spot 7 to make an intermediate sum of 10 before adding 9. In this case, the single addend 7 *complements* the current intermediate sum 3 to make a round number.

Both strategies exploit the same rationale: Two numbers, which add up to a round number are easy to add, and, when adding series of numbers, it is easier to add another number to round intermediate sums. The difference between a pair and a complement strategy is that pairs combine two external addends, whereas complements combine an internal intermediate sum with an external number.

However, both strategies come at a cost. As neither the pair nor the complement in the above example is available with adjacent elements of the linear left to right sequence, their detection requires visually searching ahead through the problem display, as well as some way of keeping track of used and skipped numbers. Thus, the use of a pair and complement strategy facilitates calculation at the expense of other resources. We hypothesize that the specific structure of this trade-off depends on the triad of factors noted above: The skill and memory capacity of the problem solver, the difficulty of the problem, and the availability of interactive resources.

Pilot Study In a pilot study we observed that the ability to rearrange or manipulate numbers on paper cards improved performance in simple addition. Furthermore, the availability of a pencil encouraged participants to use both the pair and the complement strategy, particularly with increasing problem difficulty and when numbers were presented in a 2-dimensional array. However, particular ways of using the pencil varied greatly (and included pointing, marking, copying, as well as recording intermediate sums), obstructing precise analysis of the underlying effects. Consequently, in the current study, we use computer interfaces to isolate specific interactive resources.

Method

Participants' interactions with the problem of adding numbers were operationalized as mouse actions and visual feedback on a standard computer interface. Four different *interactive modes* were distinguished:

1. *Look only*: Numbers had to be added without being able to point at them, as the mouse cursor was disabled during stimulus presentation.
2. *Point*: The mouse cursor was enabled and participants were instructed to click on numbers when adding them. When a number was clicked, a brief tone provided auditory feedback.
3. *Mark*: Mouse pointer and instructions were exactly as in 2. However, when a number was clicked, it also changed its colour from dark red to grey, thereby visually marking numbers that had been processed.
4. *Move*: Numbers could be moved on the screen using a drag-and-drop procedure.

Table 1: Examples of linear stimulus lists allowing for pairs, complements, or neither at positions x_1-x_2, y_1-y_2, and z_1-z_2.

Type	Stimulus list	Sum
Pair list:	4 3 9 7 8 6 5 4 2 1 5 9	63
Complement list:	3 1 8 6 5 3 9 4 5 7 2 9	62
Neutral list:	9 4 5 8 9 6 3 2 1 5 7 2	61
Structure:	$a\ x_1\ b\ x_2$ $c\ y_1\ d\ y_2$ $e\ z_1\ f\ z_2$	

Materials A total of 72 lists of 4, 8, or 12 single-digit numbers were generated by a Prolog program. Each list consisted of one, two, or three building blocks of the form ax_1bx_2. Three types of linear lists were distinguished: For *pair lists* x_1 and x_2 added up to 10 and the list allowed for no complements within a lookahead span of three digits. Analogically, for *complement lists* the value of x_2 plus the intermediate sum at x_1 resulted in a round number, and the list contained no pairs within a lookahead span of three digits. *Neutral lists* allowed for neither pairs nor complements within the same lookahead span. Note that none of the linear lists contained any adjacent pairs or complements (see Table 1 for some examples of stimuli). In contrast to linear lists, the elements of *spatially distributed lists* were scattered pseudo-randomly over the screen. Lists within each level of the list-length and -type factors were matched for their sums and number of possible pairs.

Stimuli presentation and data collection were controlled by a MS Windows Visual Basic program. All stimuli were displayed on a 17" computer screen using a 20pt Arial bold font of dark red colour against a white background.

Design The experiment used a mixed design, with *interactive mode* as a between-subjects manipulation and *list length* and *list type* as within-subjects factors.

Procedure Forty-four Psychology undergraduates (with a mean age of 20.3 years) took part in the experiment to receive course credit and were randomly assigned to one of the four interactive modes.

After the completion of four practice trials and a letter task to familiarize participants with their respective interactive resources, participants were instructed to add as fast as they could without making any errors. For each trial, participants pressed a button when they had added all numbers and then entered the result on another screen using a mouse-operated number pad.

Since erroneous trials were repeated at the end of the randomized sequence of trials, the experiment continued until the participant correctly added all 36 different lists (i.e. three different lists of each of three lengths and four types). On average, participants completed the experiment within 25 minutes.

Predictions The first and most basic prediction is that participants will benefit from interactive resources. In particular, following the findings of Kirsh and colleagues (1994, 1995a, 1995b), we predict that the *move* condition will elicit better performance than the *look only* condition. We also make a specific prediction concerning the comparison between the *point* and *mark* conditions. Because the latter provides an external memory for already-processed addends, it reduces the cognitive costs associated with the more sophisticated strategies of exploiting pairs and complements. Thus we predict more use of these strategies in the *mark* condition than the *point* condition, and more efficient performance as a result.

Results

Analyses of time and accuracy for the practice trials showed no differences between experimental groups at the pre-test stage. In the following report, we first focus on performance measures before considering more detailed process characteristics.

Performance

Accuracy The overall rate of errors was 13.87%. A one-way between subjects ANOVA confirmed that the number of erroneous trials in the four experimental groups differed between interactive modes [$F(3,40)=5.8$, $p=.002$, $MSE=15.2$, see Table 2 for descriptive data].

Planned comparisons revealed that participants in the *look only* condition indeed had significantly more erroneous trials than those who could *move* numbers ($t=3.28$, df=40, $p=.002$). Likewise, participants in the *point* condition made significantly more errors than those in the *mark* condition ($t=2.57$, df=40, $p=.014$). Thus, both of our specific predictions were confirmed for accuracy data.

Latency As differences in accuracy could be due to a speed-accuracy trade-off error rates and response latencies must be considered in parallel. Since the total time participants spent on the experiment is trivially longer when they made more errors we divided it by the actual number of trials for each participant to obtain an overall measure of average time per trial.

The comparison of accuracy and latency data (see Table 2) shows that no speed-accuracy trade-off contaminated the between-groups comparisons: In addition to being less accurate, participants in the *point* condition were significantly slower than those in the *mark* condition ($t=3.03$, df=40, $p=.004$), whereas the three other groups did not differ with respect to overall latency (Tukey pairwise comparisons).

Table 2: Mean number of erroneous trials and latencies per trial in seconds (and their respective standard deviations). Each interactive mode included 10 participants.

	Interactive mode			
	look only	*point*	*mark*	*move*
Errors:	8.6 (5.9)	7.8 (3.8)	3.6 (2.6)	3.2 (2.3)
Latencies:	13.8 (2.9)	17.8 (4.4)	13.3 (3.2)	12.6 (3.6)

Table 3: Mean latencies of correct solutions in seconds (a) by list length (each cell contains 132 data points) and (b) by list type (each cell contains 99 data points).

		Interactive mode			
		look only	*point*	*mark*	*move*
(a) length	4:	4.7	7.2	6.9	4.7
	8:	12.8	16.3	12.9	11.3
	12:	21.8	27.4	19.0	20.4
(b) type	pair:	13.8	18.0	12.4	11.5
	complement:	13.6	16.9	12.7	12.5
	neutral:	13.2	17.8	12.5	12.0
	spatial:	11.8	15.2	14.0	12.5

Moderating Factors

So far, we have shown that overall performance measures varied across different interactive modes. However, this standard assessment of performance based on error rates and latencies does not distinguish between different task characteristics and thus cannot uncover potential interactions between tasks and interactive resources. To analyze how performance is modulated by problem features we now qualify the global between-subjects effects by the factors of list length and type.

Accuracy The effects of list length on the frequency of errors are as expected and consistent for all interactive modes: The longer the list, the more likely participants were to add it incorrectly. Also, it made no difference to the average error rate whether a stimulus was a pair, complement, neutral, or spatially distributed list.

Since any error in calculation could effectively alter the type and length of a list, all subsequent analyses examining the effects of list length and type will be based on correct trials only.

Latency A mixed ANOVA using a 4x3x4 design was conducted to assess the effects of interactive mode, list length, and list type. Apart from significant main effects of interactive mode [$F(3,40)=4.9$, $p=.005$, $MSE=125.8$] and list length [$F(2,80)=511.6$, $p<.001$, $MSE=22.96$] it yielded significant interactions between interactive mode and list length [$F(6,80)=5.6$, $p=.001$, $MSE=23.0$] and interactive mode and list type [$F(9,120)=5.3$, $p<.001$, $MSE=6.1$].

To interpret the results of subsequent simple main effects the mean latencies are shown in Table 3. Unsurprisingly, the time needed to add a list increased for all interactive conditions as the lists' length increased (see columns of Table 3a). However, the slope of this increase

was much steeper in the *point* condition. For lists of four numbers, participants in the *look only* and *move* conditions were faster than the two other groups.

Of the eight possible simple main effects for the rows and columns of Table 3b five are significant. However, the *absence* of significant differences in three cases is more instructive: For spatially distributed lists the effects of different interactive modes levelled out [$F_{A@b4}(3,40)=2.0$, $p=.124$], which is due to the participants in the *look only* and *point* conditions being slightly faster than for other list types. Likewise, the differences between the mean latencies in the *mark* and *move* conditions for different list types failed to reach statistical significance [$F_{B@a3}(3,129)=2.3$, $p=.076$; $F_{B@a4}(3,129)=1.0$, $p=.402$], suggesting that the ability to mark and move numbers allowed participants of the corresponding groups to somehow transcend the linear and spatial constraints imposed by different list types.

Strategies

Having established that there *are* differences in performance we have to explain their genesis. We will attempt this by addressing strategy differences between groups which are reflected by features of the actual problem solving process. For this purpose, participants' cursor movements and mouse clicks in the *point*, *mark* and *move* conditions provided a rich source of fine-grained process data.

Mouse Moves per Trial When analyzing mouse cursor data, we use the term "move" to signify the physical movement from a number x_1 to a different number x_2. As each number has both value and location, moves can be characterized in terms of their distance and type, i.e., neutral, complement, pair, and triple. (In analogy to pairs, we defined a *triple* as three consecutive addends with a sum of 10.)

To obtain a measure of the amount of activity on each trial we computed the total sum of distances of all consecutive moves for each trial. A mixed 3x4x4 ANOVA on the total distance of moves per trial yielded a significant interaction between the two within-subjects factors list type and length [$F(6,180)=4.0$, $p=.010$] as well as a significant interaction between interactive mode and list length [$F(4,60)=3.5$, $p=.037$]. Whereas the first interaction merely reflects stimulus characteristics (e.g. that longer and spatially distributed lists afford longer moves) the second illustrates the modulation of moves by different interactive modes and lengths (see Table 4).

Table 4: Mean distances of cursor movements per trial in twips (1 twip = 0.05 pt = 0.01764 mm). Each cell contains 132 data points.

		Interactive mode		
		point	*mark*	*move*
length	4:	7 100	7 492	404
	8:	18 281	22 363	18 176
	12:	24 972	40 650	38 994
Total:		16 784	23 502	19 191

Table 5: Mean frequency (and standard deviations) of pairs, complements, triples, and neutral additions. Each cell summarizes data from 396 correct trials.

	Type of Move			
	Pair	*Compl.*	*Triple*	*Neutral*
point:	29.3 (14.0)	6.9 (3.2)	10.2 (7.5)	205.3 (14.7)
mark:	65.3 (38.3)	2.9 (3.2)	7.5 (2.8)	175.5 (10.3)

Whilst the increase in average move distances with longer lists was to be expected, it is notable that the slope of this increase is much steeper for the *mark* and *move* conditions. However, simple effect tests for the rows of Table 4 yielded a significant value only for lists of four numbers [$F_{A@b1}(2,30)=240.8$, $p<001$]. The lower value of the *move* condition at this length suggests why the corresponding latency was identical to the *look only* condition (see first line of Table 3a): Participants mostly chose *not* to move anything when adding short lists, but made use of their interactive potential when adding longer lists.

As the distances of moves in the *mark* and *move* conditions did not significantly exceed those in the *point* condition, activity per se cannot account for the reported differences in performance. To further illuminate potential strategy differences between groups, we have to consider process data on a within-trial level.

Choice of Next Addend At every non-last number within a stimulus participants faced the potential choice of which number to add next. We now examine the type of these choices and the corresponding move distances for the *point* and *mark* conditions, who had identical instructions (and differed only by the colour change of clicked numbers in the latter group) and both provided data on the complete paths of the chosen sequence of addends.

Table 5 contains the mean frequency of pairs, complements, triples and neutral moves chosen within correct trials. To appreciate the overall frequency of non-neutral moves we have to bear in mind that a participant had to select one or two neutral addends before being able to reach a round number and that none of the linear stimulus lists contained an adjacent pair, triple, or complement.

Pairs were the most frequent type of "rounding" move for both groups. T-tests for independent samples showed that participants in the *mark* condition chose more pairs [$t(12.6)=2.9$, $p=.012$] and fewer complements [$t(20)=2.9$, $p=.009$] than those in the *point* condition, whereas the number of triples did not significantly differ between groups [$t(14.9)=1.4$, $p=.181$].

When adjusting the frequency data for the number of *possible* pairs and complements on each particular trial, we found that participants in the *mark* condition in fact

chose to add 62.3% of all possible pairs and 2.3% of all possible complements, whereas participants in the *point* condition chose 31.4% of all possible pairs and 5.8% of the possible complements. As pair and complement strategies compete for the same addends, it is likely that the increase in complements for the *point* condition is a mere by-product of the more persistent selection of pairs in the *mark* condition.

As with the performance measures above, the differential effects of move choices were modulated by the task characteristics of list length and type. Specifically, participants in the *mark* condition predominantely pursued pairs regardless of list length and type, whereas those in the *point* group only used pairs when stimuli were short or spatially distributed.

Distance of Next Addend Additional support for the special attractiveness of pairs can be obtained when considering move distance data at the within-trial resolution.

When choosing which number to add next, participants had to balance the costs and benefits associated with the numerical value and the physical distance of each addend. If our main hypothesis about interactive problem solving applies on this micro-level, how far someone ventures in order to select a specific next number ought to vary as a function of interactive resources and number value.

Because the physical distance of moves varies trivially as a function of list length and layout, we determined how many physically closer numbers a participant skipped on each move in order to choose the next addend. By dividing the number of moves to the physically nearest unprocessed number by the total number of moves for each trial we gained a "proximity index". Its value represents the percentage of moves to the closest number per trial and ranges from 100% (indicating that the closest neighbour was always selected) to $(n\text{-}1)^{-1}$% (as at least one of the n-1 moves within a stimulus of n addends leads to a next number). The average proximity index for the *point* condition was found to be 71%, compared to a value of 61% in the *mark* condition [$t(782.3)=5.9$, $p<.001$], which indicates that marking led to a decreased likelihood to select the nearest neighbour.

To answer the question *why* participants prioritized spatially more remote addends in the other 29% or even 39% of all cases, we have to combine data on move distances and types. To quantify the price of spatial relocation a participant was willing to pay in order to make a particular type of move, we counted the number of physically closer numbers skipped for each move. Average scores of 2.20 for pairs, 1.12 for complements and 0.53 for neutral moves indicate that, to reach a pair, participants skipped about twice as many numbers than to reach a complement, whose selection still led participants to ignore about twice as many closer numbers than a neutral addend.

Moving Pairs Is there any evidence that the preference for a pair strategy generalized to the *move* condition?

Because the interactive mode of this group differed from the *look only* and *mark* conditions in that it permitted the problem solver the freedom *not* to interact, we lack the data on complete sequences of addends. However, as we observed many participants of this group either re-arranging numbers pairwise on the screen or positioning one addend of a pair physically close to the other, we computed the total distance between all possible pairs at the beginning and end of each trial. Since a mere decrease of distances between pairs could also be caused by someone moving *all* items closer together, we divided the pre- and post-trial distances between all possible pairs by the corresponding sums of distances of all possible non-pairs. A significant decrease of this ratio from initially 0.14 to 0.10 [$t(395)=12.2$, $p<.001$] allows the conclusion that pairs were moved closer towards each other than non-pairs.

Discussion

The experimental manipulation of interactive resources resulted in reliable differences in performance, which were systematically modulated by task characteristics.

Participants in the *look only* condition did well when adding short lists, but became unreliable as the number of addends increased. A similar error rate and even more pronounced increase in latencies to add long lists showed the participants in the *point* condition to be at an even greater disadvantage—presumably because they paid the additional price for interacting (clicking) without receiving the benefit of marking. As both groups had to mentally keep track of the numbers added, their strategies were more conservative and reflected specific stimulus characteristics.

In contrast, members of the *mark* and *move* groups exploited their interactive resouces to transcend the constraints imposed by stimulus and task characteristics and actively implemented a facilitative pair strategy. Their significantly faster and more reliable performance emerged as a consequence of systematic differences on a behavioural micro-level.

This finding of spontaneous adaptation to the structure of costs and benefits at the user interface supports recent attempts to describe interactive behaviour within a rational analysis framework (O'Hara and Payne, 1998; Gray and Fu, 2001). In Gray and Fu's study, a subtle increase to the cost of external information (an eye-movement or a single mouse-click) led to users of a simulated VCR relying on imperfect memory. In our study a relatively subtle change to the information display reduces internal memory load and thus enables a more sophisticated strategy for ordering addends. What is important about studies like these is not so much that small changes to the task environment can produce reliable shifts in behaviour but that an analysis of the interactions between physical and cognitve costs and benefits can predict and explain the particular behaviours that emerge.

In the current experiment, the additional resources provided by the more powerful interactive modes were all available relative cheaply (as are the so called

"epistemic" and "complementary actions"—like rotating objects or moving coins—in the studies by Kirsh, as cited above). What happens when complementary actions become more expensive, in terms of time or mental effort? Even in the current study, the results from the *move* condition suggest that disuse of interactive resources can sometimes be adaptive. In future work, we propose to investigate such questions by directly manipulating costs, following the methodology of O'Hara and Payne (1998), and by asking participants to explicitly choose between modes of interaction, using the choice-no-choice paradigm of Siegler and Lemaire (1997).

In this experiment, participants spontaneously, and almost instantly adopted a strategy which was tuned to their interactive resources. This contrasts with findings that people often are very inflexible in their behavioural routines, and continue to use dysfunctional strategies even when more efficient alternatives are available (Carroll & Rosson, 1987). To address this apparent discrepancy between rapid adaptation and rigid perseverance future studies will have to incorporate issues of learning and transfer.

The implications of this line of research are manifold:

On a theoretical and conceptual level, a strong version of the interactive perspective challenges the distinction between agent and environment, and promises to bridge the gap between cognition and action (Clark, 1997; Kirsh, 1996).

Methodologically, the dynamic interplay of factors illustrates that studies of interactive cognition have to strive for a very fine-grained resolution. To study the features of an agent, task, or task environment in isolation would fail to capture the multi-faceted nature of effects and misrepresent the complex balancing act of successful problem solving.

Finally, the study of interactive problem solving promises practical applications. Several studies have now shown that subtle changes in interactional resources can lead to substantial differences in performance. The challenge for interface design is to understand the complex structure of costs and benefits imposed by different environments, and to use this understanding to produce information displays that encourage effective interactions.

Acknowledgments

We would like to thank Suzanne Charman, Will Reader, and the anonymous reviewers for their comments and suggestions on an earlier draft. This research was supported by ESRC Research Studentship Award No. R00429934325 to HN.

References

Carroll, J.M., & Rosson, M.B. (1987). The paradox of the active user. In J.M. Carroll (Ed.), *Interfacing thought: Cognitive aspects of human-computer interaction*. Cambridge, MA: The MIT Press.

Clark, A. (1997). *Being there: Putting brain, body, and world together again*. Cambridge, MA: The MIT Press.

Gray, W.D. & Fu, W.-t. (2001). Ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head: Implications of rational analysis for interface design. In *CHI Letters, 3(1)*.

Kirsh D. (1995a). Complementary Strategies: Why we use our hands when we think. In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kirsh, D. (1995b). The Intelligent Use of Space. *Artificial Intelligence*, *73*, 31–68.

Kirsh, D. (1996). Adapting the environment instead of oneself. *Adaptive Behavior*, *4*, 415–452.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, *18*, 513–549.

Larkin, J.H. & Simon, H.A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11*, 65–99.

Maglio, P.P., Matlock, T., Raphaely, D., Chernicky, B., & Kirsh D. (1999). Interactive Skill in Scrabble. In *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society* (pp. 326–330). Mahwah, NJ: Lawrence Erlbaum Associates.

Mayer, R.E. (1992). *Thinking, problem solving, cognition*. New York: W.H. Freeman.

Miller, G.A., Galanter, E., & Pribram, K. (1960). *Plans and the structure of behavior.* New York: Holt, Rinehart and Winston.

Monk, A. (1998). Cyclic interaction: a unitary approach to intention, action and the environment. *Cognition, 68*, 95–110.

Newell, A., & Simon, H.A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

O'Hara, K.P., & Payne, S.J. (1998). The effects of operator implementation cost on planfulness of problem solving and learning. *Cognitive Psychology*, *35*, 34–70.

Payne, S.J. (1991). Display-based action at the user interface. *International Journal of Man-Machine Studies, 35*, 275–289.

Siegler, R.S., & Lemaire, P. (1997). Older and younger adults' strategy choices in multiplication: Testing predictions of ASCM via the choice/no-choice method. *Journal of Experimental Psychology: General*, *126*, 71–92.

Zhang, J. (1997). The nature of external representations in problem solving. *Cognitive Science*, *21*, 179–217.

Zhang, J. & Norman, D.A. (1994). Representations in distributed cognitive tasks. *Cognitive Science*, *18*, 87–122.

On the Normativity of Failing to Recall Valid Advice

David C. Noelle (NOELLE@CNBC.CMU.EDU)
Center for the Neural Basis of Cognition; Carnegie Mellon University
Pittsburgh, PA 15213 USA

Abstract

Instructed category learning tasks involve the acquisition of a categorization skill from two sources of information: explicit rules provided by a knowledgeable teacher and experience with a collection of labeled examples. Studies of human performance on such tasks have shown that practice categorizing a collection of training examples can actually interfere with the proper application of explicitly provided rules to novel items. In this paper, the normativity of such exemplar-based interference is assessed by confronting a model of optimal memory performance with such a task and comparing the "rational" model's behavior with that exhibited by human learners. When augmented with a rehearsal mechanism, this optimal memory model is shown to match human responding, producing exemplar-based interference by relying on memories of similar training set exemplars to categorize a novel item, in favor of recalling rule instructions.

Introduction

Contemporary studies of human category learning have tended to focus on the acquisition of general knowledge about a new concept exclusively from exposure to a collection of labeled examples. In common learning environments, however, students attempting to learn a categorization skill are frequently provided with more than a set of training examples. In particular, learners are often explicitly instructed in the nature of a new category before being presented with instances. They are provided with definitional sentences and explicit rules (e.g., "an equilateral triangle has at least two sides of the same length" or "bugs with six legs are insects"). Direct instruction of this kind can rapidly provide a basic understanding of a new category, while experience with examples can further shape and refine that initial understanding (Klahr and Simon, 1999).

While it is common for the process of explicit instruction following and the process of induction from examples to cooperate to produce quick and robust learning, there are situations in which these two learning processes actually compete. Specifically, practice at classifying a set of training examples can cause learners to violate explicitly provided categorization rules when classifying novel items. Extensive experience with examples can lead learners to categorize novel instances according to similarity to training items, rather than according to categorization rules communicated through explicit instruction. Thus, novel items which are highly similar to training examples from another category come to be misclassified as a result of practice.

This exemplar-based interference effect, in which experience with examples interferes with proper instruction following, was investigated by Allen and Brooks (1991), as well as others (Brooks et al., 1991; Neal et al., 1995; Noelle and Cottrell, 2000). Such interference in category learning is mirrored by similar difficulties in a wide variety of learning contexts, such as when students come to solve math or science problems by analogy to previously seen problems, rather than by application of formal principles and techniques communicated through direct instruction. Learners appear to have a tendency to disregard perfectly valid explicit advice in favor of knowledge induced from experiences with examples.

Exemplar-based interference might be seen as the result of limitations of the cognitive system, such as imperfect working memory efficacy (Noelle and Cottrell, 2000) or difficulties recalling and applying abstract, linguistically encoded, rules. There is another alternative, however. It is possible that human learners neglect explicit instructions in favor of experienced exemplar-similarity information because the latter form of information tends to be more reliable in a wide variety of learning contexts. Exemplar-based interference may be the result of an essentially normative process of weighting sources of category information according to the previously established utilities of those sources.

There are many aspects of common learning situations which may encourage students to rely more heavily on examples than on explicit rules. Consider, for example, how the instructions provided by teachers are frequently approximate and heuristic. Advice is often implicitly limited to a particular range of circumstances, and there are often exceptions, even within this range, to explicitly provided rules. Also, teachers are sometimes in error. In short, human learners may have strong reasons to doubt the perfect accuracy of offered categorization rules. In comparison, exemplar similarity may be seen as a highly reliable indicator of category membership. Most categories, after all, involve clusters of similar objects, suggesting that similarity might be the best tool for predicting the category labels of novel instances.

Even if considerations of teacher reliability are ignored, there are other rational reasons for a learner to

rely preferentially on training experiences. In general, recalling past experiences with features similar to those of the current situation is often more useful than recalling dissimilar experiences. Thus, when faced with the task of categorizing a novel stimulus item, learners may be naturally inclined to recall other similar items rather than an explicit rule, which, due to its linguistic encoding, may bear little surface similarity to the situation at hand. Also, the recollection of experiences which are recent and frequently recurring is, on average, more useful when facing a novel challenge than recalling rare experiences from one's distant past. Thus, when performing an instructed category learning task, it may be reasonable for a learner to selectively recall the training items which were recently and repeatedly studied in favor of a briefly presented rule. In short, we may conjecture that exemplar-based interference arises from a rational tendency to rely on similar, recent, and frequent past experiences when faced with a novel situation.

In order to evaluate this conjecture, this paper reports on the modeling of the exemplar-based interference results of Allen and Brooks (1991) using the normative, or "rational", account of memory formulated by Anderson (1990). The goal is to investigate the degree to which exemplar-based interference can be explained in terms of a Bayes optimal learning process, given some assumptions about the common demands placed on human memory. The human performance results are reviewed first, followed by a description of Anderson's optimal memory model. The results of applying the model to this domain are then presented.

Human Performance

Allen and Brooks (1991) performed a number of experiments demonstrating the way in which experience with labeled training exemplars can interfere with instructed rule following. In their Experiment 1, learners were asked to categorize cartoon illustrations of fictional animals into one of two categories, based on how the animals were said to construct their homes: the "builders" and the "diggers". The appropriate category for each animal was strictly determined by it's physical features. Each animal was composed of specific selections for five binary attributes: angular body shape or rounded body shape, spots or no spots, short legs or long legs, short neck or long neck, and two legs or four legs. Only three of these attributes were ever relevant for classification, however: body shape, presence or absence of spots, and leg length. The animals were always depicted against color backgrounds, displaying four different outdoor environments. From this space of $2^5 \times 4 = 128$ different possible stimuli, only 16 were actually used. These 16 items were carefully chosen to include two animals with each possible level of the three relevant attributes. The irrelevant features were selected so that each stimulus item would have exactly one "partner" item — an item which differed from it only in the presence or absence of spots. Otherwise, each animal differed from each other animal in at least two attributes.

Experimental participants were provided with explicit categorization rules for discerning the "builders" from the "diggers". These always took the form of "2 of 3" rules, in which a target category was described as all animals with at least two of a list of three features (e.g., builders have two or more of the following features: angular body shape, spots, long legs). The rules were carefully chosen so that the 16 stimuli were equally split between the two categories. Also, the exemplars were partitioned into a training set and a testing set so that no two "partnered" items were in the same set. This resulted in exactly half of the testing set items having their partner items in the opposite category. These testing items were the ones for which interference was predicted.

The learners were presented with a training phase which consisted of seeing each of the 8 training set items five times, presented in a random order, for a total of 40 trials. When a stimulus image appeared on the screen, learners were to categorize it as quickly as possible, without sacrificing accuracy. Then, a sequence of two slides would be shown, illustrating how the animal actually constructed its home, identifying it as a builder or a digger. A subsequent testing phase involved soliciting categorization responses from the participants without providing any form of feedback on their decisions. During this testing phase, each training set stimulus was presented 4 times and each testing set stimulus was presented once, for a total of 40 testing trials.

There were two main results of this experiment. First, accuracy on the items whose "partners" were in the opposite category was much worse than on the other testing set items — around 55% correct as compared to 80%. This was a strong indication of exemplar-based interference. Second, the response time for correctly classified items was much larger for items whose "partners" were in the opposite category. This was interpreted as extra caution on the part of the learners when facing these "tricky" stimulus items. In other words, even when exemplar-based interference did not cause error, it at least caused a slowing of behavior.

Allen and Brooks argued that explicit memories for individual stimulus items played an important role in the production of this interference effect. The presentation of a testing set stimulus was seen as provoking a recollection of that item's "partner" in the training set, with the category label of that training set item often being assigned to the new stimulus in lieu of a label based on explicit rule application. Following this intuition concerning the centrality of memory to this effect, we have attempted to model these data using a previously explicated account of optimal memory performance.

Anderson's Rational Memory

The hypothesis explored here is that the behavior of the learners examined by Allen and Brooks can be characterized as normative — as the natural result of employing a memory system which is optimal in a Bayesian sense. This raises the question of how an optimal memory system would respond in this domain. Anderson and Mil-

son (1989) have proposed a "rational" model of memory which might be employed to address this question.

Initially, one may think that an optimal memory is a perfect memory. *Everything* is to be stored in every detail, without degradation, for an unlimited amount of time. This overlooks one very important function of memory, however, and that is to recall only those memories which are relevant to the current task. Without this ability of selective recall, a memory is essentially useless, even if (or especially if) it contains every detail that was ever experienced. Thus, the task faced by an optimal memory is the identification of those memory traces which would be most useful in the current situation.

In Bayesian terms, the goal is to determine, for each memory trace, the probability that that trace would be useful in the current situation. In Anderson's model, this is called the "need probability" of a trace. An optimal memory is seen as one which retrieves exactly those traces with the highest need probabilities in the current context. The question then becomes one of calculating the need probability for each memory trace.

In this model, the need probability is seen as a function of two components: the *desirability* of the trace and the *association* between the trace and the current context. The desirability of a memory trace is a measure of the average utility of the trace — a kind of base rate of appropriateness. The desirability of a trace is to be induced from its history of use. Recent and frequent retrieval of a memory trace is indicative of high desirability. The association between the trace and the current context is a kind of normalized likelihood of the context given that the trace is needed. This term increases the need probability with increased similarity between the context and the trace. Both of these components of the need probability are seen as normative properties of the situation, unbiased by predispositions of the agent. In brief, the optimal memory system computes the need probability of each memory trace, conditioned on the current context and on the history of past retrievals of that trace.

Mathematically, if A represents the event that a given memory trace is needed in the current context, H_A represents the complete retrieval history of that trace, and Q is the current context, then the conditional need probability is $P(A|H_A \ \& \ Q)$, which may be decomposed as follows:

$$P(A|H_A \ \& \ Q) \ = \ P(A|H_A) \times \frac{P(Q|A)}{P(Q)}$$

Note that this assumes that Q and H_A are both independent and conditionally independent with respect to A. If Q is taken to be composed of a collection of mutually independent features, then this expression may be written as:

$$P(A|H_A \ \& \ Q) \ = \ P(A|H_A) \times \prod_{i \in Q} \frac{P(i|A)}{P(i)}$$

This formulation allows for the separate calculation of a *history factor*, $P(A|H_A)$, and a *context factor* which measures the association between the memory trace and each feature of the current context, $P(i|A)$.

The calculation of the history factor requires some assumptions about the desirability of memory traces. Each trace is taken to start at some desirability level, λ_0, when it is first generated. Over the range of memory traces, these initial desirabilities are assumed to have a gamma distribution with parameter b and index ν. This means that no traces have an initial desirability of zero, most have some small initial desirability, and a very few have a high value for this variable. Furthermore, desirability is assumed to decay exponentially over time, with a decay rate of δ, where this rate of decay varies over the traces. It is assumed that δ is exponentially distributed with parameter α. Together, these assumptions paint a picture of memory traces with various initial desirabilities, decaying exponentially over time at various rates. Some memory traces start out with a high desirability and decay only slowly, like, say, the trace for your own name. Other traces start out with a low probability of use, like instructions on how to help a heart attack victim, but the desirability does not decay much with time. Some memories are very important but only for a short time, such as the memory for how much money was handed to a cashier before receiving change. Most trivia start out with a low desirability and decay rapidly.

One phenomenon not captured by this characterization is the way in which certain memory traces might become very useful again, after a long period of unimportance. To remedy this oversight, it is assumed that memory traces occasionally experience "revivals", at which time their desirabilities are returned to their original levels. The probability of a revival of a memory trace is assumed to decay exponentially with the time since the trace's introduction, with rate β.

This formulation provides a characterization of the probability distribution of possible trajectories of desirability over time. Recall, however, that what is needed is the distribution of histories of actual trace retrievals:

$$P(A|H_A) \ = \ \frac{P(A \ \& \ H_A)}{P(H_A)}$$

If we assume that a trace is retrieved with a probability proportional to its desirability, we can compute $P(H_A)$ by integrating over all possible values of initial desirability, decay rate, and revival history. This value is:

$$P(H_A) \ = \ \int\int \ P(H_A|\delta \ \& \ R) \ p(\delta) \ p(R) \ \ d\delta \, dR$$

where δ is a decay rate and R is a particular revival history. Note that, in this expression, the initial desirability has already been integrated over. The main term in this double integration has the form:

$$P(H_A|\delta \ \& \ R) \ = \ \frac{b^{\nu}(n+\nu-1)!}{D^{n+\nu-1}(\nu-1)!} \ \prod_{i=1}^{n} e^{-\delta(H_i - r_i)}$$

where n is the number of retrievals in H_A, H_i is the time of the *ith* retrieval, r_i is the time of the revival which most

immediately preceded the *ith* retrieval, and D is:

$$D = b + \frac{1}{\delta} \sum_{j=0}^{m} \left(1 - e^{-\delta(R_{j+1} - R_j)}\right)$$

where m is the number of revivals, and R_j is the time of the *jth* revival. All other variables in these expressions are parameters from the previously discussed probability distributions. In short, an expression for the value of $P(H_A)$ is available in the form of the double integral above.[1] This double integral ranges over an infinite space of δ values and possible revival histories. In order to estimate the value of this expression, a Monte Carlo integration may be performed, sampling decay rates and revival histories from their respective distributions. In this way, an estimate of $P(H_A)$ can be calculated.

Note that $P(A \ \& \ H_A)$ can be calculated in exactly the same fashion as $P(H_A)$ simple by including an additional retrieval of the memory trace at the current moment. As previously noted, the ratio of these two probabilities is the needed history factor, $P(A|H_A)$.

The calculation of the context factor is much easier to perform, mostly due to some simplifying assumptions. To compute the contribution of the association between the trace and the current context, it is assumed that the trace is composed of features which contribute independently to the need probability of the trace. These features are assumed to be mutually independent, even when conditioned on any feature of the current context. Thus, the context factor can be written as:

$$\prod_{i \in Q} \frac{P(i|A)}{P(i)} = \prod_{i \in Q} \frac{P(A|i)}{P(A)} = \prod_{i \in Q} \prod_{x \in A} \frac{P(x|i)}{P(x)}$$

All that remains is to determine the associative strengths between features of the current context and features of the memory trace, expressed as $P(x|i)$, which may be selected in a manner sensitive to the specific stimuli used.

Anderson and Milson (1989) showed that this optimal memory model matched human performance in many ways. This calculation of the probability of retrieval was found to predict recency and frequency effects, and the model was shown to be consistent with effects arising from varying the temporal spacing between the presentations of stimuli. This complex retrieval probability computation accounted for effects of word frequency on the memorization of word lists, priming effects, and various fan effects. Most all of these calculations were performed with fixed values for the distribution parameters: $b = 100$, $\nu = 2$, $\alpha = 2.5$, and $\beta = 0.04$.

Modeling Exemplar-Based Interference

Following the theorizing of Allen and Brooks (1991), their instructed category learning task can be viewed as a memory task. When initially given the explicit rule for categorizing the fictional animals, the learner must remember this rule, and it must be recalled when it is needed to categorize a stimulus item. The rule need not always be recalled, however, as it will be sufficient in many cases to simply remember a previous presentation of the specific stimulus being viewed and its corresponding category label. This characterization of the task makes Anderson's rational memory model applicable to an optimality analysis of instructed category learning.

A computer program was written which simulated the performance of Anderson's rational memory on the experimental task examined by Allen and Brooks (1991). Initial instruction involved the creation of a memory trace for the given categorization rule, and the retrieval of that trace for ten consecutive time steps, representing a study period. After this instruction period, the training set items were presented to the optimal memory, one at a time, in the same manner as they were presented to human participants. With each presentation, the need probability of each existing memory trace was estimated in the context of the current stimulus. The memory trace with the highest need probability among those traces that contained a category label was retrieved from the memory.[2] The category label of the retrieved memory trace was taken to be the response provided by the optimal memory system to the current stimulus. Note that the memory trace for the explicit rule was seen as containing the correct category label for every stimulus item.

During the training phase, the solicitation of a categorization judgment from the memory was followed by the incorporation of performance feedback information. The memory system responded to feedback by immediately retrieving the memory trace corresponding to the current stimulus, or, if this was the first presentation of the given item, by generating and retrieving a new trace for the stimulus, marked with the given category label.

After the training phase, the optimal memory experienced a testing phase equivalent to that presented to the human learners, involving a mix of training set items and new testing set items. The protocol for memory trace retrieval during the testing phase was the same as during training, except that none of the newly generated memory traces contained category label information, as no feedback was provided to the humans during this phase. Categorization errors made by the memory system during the testing phase were examined for signs of exemplar-based interference: relatively poor accuracy on those testing set stimuli whose "partner" items in the training set were in the opposing category.

To calculate the history factor of the need probabilities, the same parameters that were used by Anderson and Milson (1989) were used in this simulation: $b = 100$,

[1] Note that this expression is different than that provided in the appendix of Anderson and Milson (1989). When this error was brought to the attention of the authors, they provided the software that they had used to perform their calculations. It was discovered that the error was only in their appendix and not in their software.

[2] During the testing phase it was possible that the memory trace with the highest need probability would be a memory of a previous presentation of an unlabeled item. Such a memory would not be of much use for making a categorization judgment. Thus, this retrieval was restricted only to those memory traces which contained explicit category information.

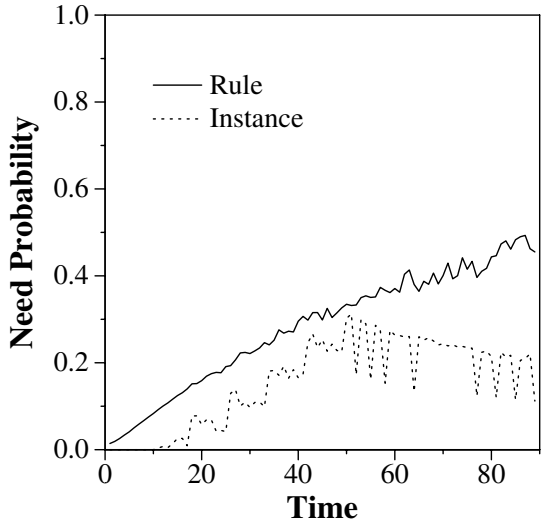


Figure 1: Results from the Optimal Memory Model: The need probability of the rule memory trace is plotted against the maximum need probability among the traces for the training set items. Note that the training phase ran from time step 11 through 50, and the testing phase ran from time 51 through 90.

$\nu = 2$, $\alpha = 2.5$, and $\beta = 0.04$. To calculate the context factor, the presentation of a stimulus was seen as providing a context consisting of 5 binary features (i.e., the attributes of the fictional animals) and one 4-ary feature (i.e., the background). Memory traces were seen as containing these six features, plus an optional category label. The associational strength between context and trace features was taken to be $P(x|i) = 0.65$. These features were taken to be pictorial in nature, so the memory trace for the explicit verbal rule contained none of these features. The Monte Carlo integration process employed by the optimal memory model consistently used $100,000$ samples in the calculation of each need probability estimate.

A summary of the results of this computation are shown in Figure 1. Plotted in that graph is the calculated need probability of the explicit rule memory trace and the highest need probability over the training set exemplar traces, both over time. Note that the training phase began at time step 11 and ended at time step 50, and the testing phase ran from time step 51 through time step 90. The primary result shown in this graph is that the rule always dominated over the exemplars. This meant that the rule was always retrieved in preference to traces for previously viewed items. In other words, the optimal memory produced perfect rule following behavior with no sign of interference. Even when the optimal memory system was modified to stochastically retrieve traces in a manner proportional to their need probabilities (rather than always retrieving the trace with the highest need probability), errors on stimulus items with "partners" in the opposite category averaged only 12%, as compared to the 45% error exhibited by humans.

These results were found, however, to be very sensitive to the associational strength that was used, $P(x|i)$. If this value was substantially increased above 0.65, then the memories for the training set items would immediately and persistently dominate over the trace for the rule. Under such higher settings of the associational strength, the optimal memory model would produce interference during the appropriate portions of the testing phase, but it would not produce expected behavior early in the session. In particular, the explicit rule would almost never be used. In short, this initial simulation of the optimal memory model of instructed category learning did not match human performance very well at all.

Anderson had some similar problems with his rational memory model when he compared its performance to human behavior (Anderson, 1990). While human responding matched his rational memory calculations in a number of domains, there were some aspects of human performance which could only be fit by the model with the help of an additional assumption. This assumption was that the system would covertly rehearse recently retrieved traces. He added to the memory model a rehearsal buffer which contained the 4 most recently retrieved memory traces. On each time step, each trace in the rehearsal buffer had a 0.2 probability of being rehearsed on that time step. Rehearsal simply involved the retrieval of that trace from memory. Increasing the number of retrievals of a trace through rehearsal would expand its retrieval history, H_A, and would thereby increase the history factor, $P(A|H_A)$, for that trace. Anderson added this rehearsal strategy, admitting that it stepped beyond the bounds of an optimality analysis. Still, such an augmented analysis was considered worthwhile, since it could show that human performance is optimal up to the inclusion of such rehearsal strategies. Indeed, that was exactly what Anderson demonstrated for a number of memory phenomena.

Following Anderson's lead, the optimal instructed category learning simulation was augmented with a 4 element rehearsal buffer. As in Anderson's work, the probability of rehearsal for each item in the buffer was set to 0.2 per time step. The memory trace for the instructed rule was allowed to occupy the buffer and be rehearsed, just like any other memory trace. The associational strength parameter was kept at 0.65.

Adding this rehearsal mechanism had a substantial impact on the behavior of the optimal memory, as shown in Figure 2. With rehearsal, the explicit rule maintained its perceived utility through much of the training phase, but was overcome by exemplar similarity by the time the testing items were presented. This produced consistent errors on those stimuli whose "partners" were in the opposite category. When traces were retrieved stochastically, in proportion to their need probabilities, the frequency of error on such items was 42%, comparing favorably to the 45% error exhibited by human learners. Thus, the rational memory model, when augmented with rehearsal, appears to be consistent with the observed in-

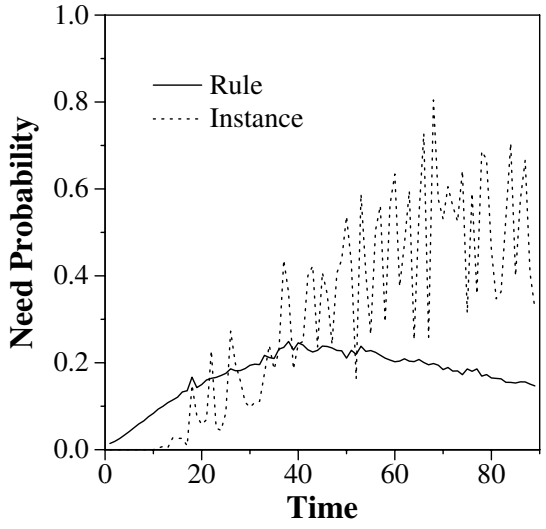


Figure 2: Results from the Optimal Memory Model With Rehearsal: Once again, the need probability of the rule memory trace is plotted against the maximum need probability among the traces for the training set items.

terference effect in instructed category learning.

Discussion

In many situations, it is more useful to remember a highly similar episode from the past than to recall generally applicable instructions. The rational memory model of Anderson and Milson (1989) is a formalization of the process of optimally predicting when such a situation has arisen. The unaugmented optimal memory model specifies that, within the experimental design of Allen and Brooks (1991), the explicit rule should almost always be preferred if similarity is not very predictive (i.e., when the associational strength is low), and a memory for specific instances should almost always be preferred if similarity is sufficiently predictive (i.e., when the associational strength is high). This is not consistent with human performance, however, where errors on "tricky" testing set items appeared only 45% of the time.

However, if the rational memory model is augmented with a rehearsal mechanism, as is needed to explain performance on other memory tasks (Anderson, 1990), the resulting need probabilities match human performance much more accurately. This suggests that the interference effect of interest may arise in the interaction between an optimal memory mechanism and a rehearsal strategy. One prediction of this calculation is that experimental manipulations which hinder rehearsal will reduce exemplar-based interference.

Note that, in these simulations, the memory trace for the explicit rule shared no features with the stimulus presentation contexts. This was intended to model the fact that the stimuli were pictorial, while the rule was linguistic. In fact, if the features itemized in the explicit rule are associated with the corresponding stimulus features with the same associational strength as used elsewhere in these simulations (0.65), the explicit rule comes to dominate over exemplar memory traces, even in the augmented model. It is a surprising fact is that this property of the model actually reflects human responding. Exemplar-based interference virtually disappeared when Allen and Brooks (1991) presented the animal stimuli not as pictures but as word lists — allowing the stimulus features and the explicit rule terms to literally match.

In summary, while this analysis does not rule out other potential explanations of exemplar-based interference, it offers the tantalizing possibility that the human tendency to ignore explicit instructions in favor of information provided by example experiences may be essentially adaptive when considered within the context of the common demands placed on the cognitive systems responsible for learning and memory.

Acknowledgements

This work was supported, in part, by a National Research Service Award (# 1 F32 MH11957-01) from the USA National Institute of Mental Health. Thanks are extended to Garrison W. Cottrell, Richard Anderson, and two anonymous reviewers for their helpful comments.

References

Allen, S. W. and Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, 120(1):3–19.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Studies in Cognition. Lawrence Erlbaum, Hillsdale, New Jersey.

Anderson, J. R. and Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, 96(4):703–719.

Brooks, L. R., Norman, G. R., and Allen, S. W. (1991). The role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General*, 120(3):278–287.

Klahr, D. and Simon, H. A. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. *Psychological Bulletin*, 125(5):524–543.

Neal, A., Hesketh, B., and Andrews, S. (1995). Instance-based categorization: Automatic versus intentional forms of retrieval. *Memory & Cognition*, 23(2):227–242.

Noelle, D. C. and Cottrell, G. W. (2000). Individual differences in exemplar-based interference during instructed category learning. In Gleitman, L. R. and Joshi, A. K., editors, *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*, pages 358–363, Philadelphia. Lawrence Erlbaum.

How is Abstract, Generative Knowledge Acquired? A Comparison of Three Learning Scenarios

Timothy J. Nokes (**tnokes@uic.edu**)
Stellan Ohlsson (**stellan@uic.edu**)

Department of Psychology
The University of Illinois at Chicago
1007 West Harrison Street (M/C 285)
Chicago, IL 60607, U.S.A.

Abstract

Several theories of learning have been proposed to account for the acquisition of abstract, generative knowledge including schema theory, analogical learning and implicit learning. However, past research has not compared these three theories directly. In the present studies we instantiated each theory as a learning scenario (i.e., direct instruction, analogy training and implicit training) and then tested all three training groups on a common problem. Results show that the analogy training groups and one of the direct instruction groups performed significantly better than the other groups on problem solving performance. The findings are interpreted in terms of opportunity to practice generating a response of the relevant type.

Theories of Deep Learning

In order to solve complex, novel problems one must be able to retrieve previously learned information from memory and apply it to the current situation. For instance, students learning geometry need to be able to apply mathematical formulas acquired during study to novel problems encountered at test. Although surface features of the problems change (e.g., specific values: a=5 to a=15) the abstract operators used to solve the problems stay the same (e.g., the Pythagorean theorem: $a^2 + b^2 = c^2$). Thus, in order for the knowledge gained from study to be helpful on the test it must be both abstract and generative. How such deep knowledge is acquired remains a central question for researchers in psychology, philosophy and education.

Several theoretical explanations have been proposed as to the origin of such abstract, generative knowledge including: schema theory (Marshall, 1995; Thorndike, 1984), analogical learning (Gentner, 1983; Holyoak & Thagard, 1988) and implicit learning (Reber, 1989).

Research on schema theory has shown that abstract knowledge is constructed during various types of higher-order cognitive activities including text comprehension (e.g., Kintsch & van Dijk, 1978; Thorndike, 1977), problem solving (e.g., Marshall, 1995) and direct instruction (e.g., Ohlsson & Hemmerich, 1999; Ohlsson & Regan, in press). For the purposes of this paper we are not concerned with the induction hypothesis of schema acquisition but instead with whether a schema can be taught directly (Ohlsson & Hemmerich, 1999). Although schema theory has provided much insight into the nature and form of abstract knowledge representations (e.g., Bobrow & Collins, 1975) it has done little to articulate a specific theory for how abstract schemas are acquired.

A second major theoretical proposal is the analogical learning hypothesis. Research on analogical learning suggests that one acquires deep knowledge through a systematic process in which a person retrieves an analog from memory and maps the underlying conceptual structure to a novel problem (Gentner, 1983; Holyoak & Thagard, 1988). In a typical analogical learning experiment participants first solve a source problem (e.g., story problems: Gick & Holyoak, 1983; or algebra problems: Reed, 1987) and then solve a test problem that has different surface features (i.e., a different context) but retains the deep relational features of the source problem. When participants are given the hint to use the source problem to solve the test problem they perform better than a control group who did not receive prior training, indicating that explicit knowledge of the prior solution procedure facilitates subsequent problem solving.

In contrast to the prior two theories research on implicit learning suggests that knowledge acquisition is a passive, inductive process that is independent of any intention to learn (Reber, 1989; Seger, 1994). In the training phase of artificial grammar learning – a typical implicit learning paradigm – the participants memorize letter strings that are generated from an artificial grammar. Participants are not informed of the rule-based nature of the memorization strings until after the training phase. In the test phase, the participants are given a classification task in which they are asked to judge whether or not new letter strings, half generated by the relevant grammar and half violating one or more of the rules, are like those memorized during the training phase. A large amount of evidence (Reber 1989; Seger; 1994; Stadler & Frensch, 1998) shows that participants perform better than chance in the test phase, indicating that they have acquired some knowledge of the underlying grammar.

These three theories present a complicated if not contradictory picture of knowledge acquisition. Each theory has a history of empirical support, experimental paradigms and explanatory problems associated with it. In order to

further explore and understand the relationship between these theories comparative empirical studies are essential. In the present studies we examine all three theories via a single experimental paradigm. This allows us to compare and examine the type of the knowledge generated by the various learning scenarios associated with each theory. How does that knowledge function on subsequent tasks? Is the knowledge representation abstract? Is it generative?

We instantiate each theory as a different type of learning scenario (i.e., direct instruction, analogical training and implicit training) and then test all three training groups on a common problem, Thurstone's sequence extrapolation task (Thurstone & Thurstone, 1941).

Sequence extrapolation problems have been studied from a cognitive perspective by Kotovsky and Simon (1973) and Simon (1972), among others. In this type of problem, the problem solver is given a sequence of symbols (usually letters) generated in accordance with some complex pattern and asked to extrapolate it. In order to solve the problem, he or she must first *discover* the pattern in the given segment of the symbol sequence and then articulate that pattern to *generate* the next N positions of the sequence.

In the current studies, experiments were divided into a training phase and a test phase. Participants in the implicit training condition memorized letter strings that had the same abstract pattern as that used in the test phase. Participants in the analogy training condition solved source problems that had the same abstract pattern as that used in the test problems. Participants in the instruction condition read a general tutorial on how to solve sequence extrapolation problems and studied the abstract rules of the pattern for each target problem.

In the test phase of the experiments all participants solved two types of extrapolation problems, a target problem and a transfer problem. The *target problem* had the same deep structure (i.e., the relations between the elements of the pattern) as that used in the sequences of the training procedures. The *transfer problem* was generated from the target problems by 'stretching' relations between letters, e.g., "forward 1 step in the alphabet" becomes "forward 2 steps" and "backwards 1" becomes "backwards 2" (see Table 1).

The goal of the present studies is to examine the nature and function of the knowledge created by each of the three training scenarios. If the knowledge gained is generative (i.e., can articulate a sequence of temporally related actions) and abstract (i.e., not bound to surface features) then performance should be facilitated on the target task. If the knowledge is of a higher level of abstraction it should facilitate problem solving performance on the transfer problem.

Experiment 1: Analogy vs. Implicit Training

Experiment 1 compares the effect of different types of analogical training (i.e., solving 1 problem three times vs. solving 3 structurally similar problems once) to different levels of implicit training (i.e., memorizing 6 vs. 18 instances of the pattern) on problem solving task performance.

Method

Participants One hundred and twenty seven undergraduate students from the University of Illinois at Chicago participated in return for course credit.

Materials The target tasks were two sequence extrapolation problems with a periodicity of six items for problem 1 and seven items for problem 2. Each task was instantiated as both a target and transfer problem; see Table 1. Target and transfer problems were related in that they contain similar over-arching pattern types but differed in the particular instantiation of the relations (i.e., manipulating some of the relations of the target pattern by a magnitude of 2 for transfer problems). To enable participants to detect the pattern, the given segments were 12 items long for task 1 and 14 items long for task 2. That is, they covered two complete periods of the underlying pattern. The extrapolation problems were created specifically for this experiment with a design similar to the problems used by Kotovsky and Simon (1973).

Table 1. Two sequence extrapolation problems and their associated transfer problems.

Problem Type	Given letter or number sequence & the correct 8-step extrapolation
Problem 1	
Target	E F D G C O F G E H D P G H F I E Q H I
Transfer	E G D I C O G I F K E P I K H M G Q K M
Problem 2	
Target	A C Z D B Y Y D F X G E W W G I V J H U U J
Transfer	A E Z G C X X G K V M I T T M Q R S O P P S

There were also three extrapolation *training* problems for each target task. The three training problems followed the exact same pattern as the associated target problem; see Table 2a for an example. Training problems were constructed so they do not overlap (i.e., do not share any of the surface features) with each other or the target problems. The single analog group was trained on the first of the three training problems and the multiple analog group was trained on all three.

In addition, there were 36 letter training strings consisting of 12 letters for task 1 and 14 letters for task 2, eighteen strings for each problem. The eighteen strings

associated with each problem followed the exact same pattern as the given sequence for that problem; see Table 2b for an example. The low implicit participants were trained on six strings per task and the high implicit participants were trained on eighteen strings per task.

Table 2. Two training sequences for Problem 1.

Example	
A. Source Problem:	I J H K G S J K I L H T
B. Training String:	M N L O K W N O M P L X

Each participant was tested on a Macintosh computer with a 14' color monitor, standard keyboard and mouse. All stimuli were presented in black 30pt font in the center of the screen. The experiment was designed and generated using PsyScope software. Target and transfer problems and associated training stimuli were counter-balanced across all conditions.

Design and procedure The participants were randomly assigned to one of four groups: *single analog* (n = 25), *multiple analog* (n = 23), *low implicit* (n = 26), *high implicit* (n = 23). In addition, a separate *control group* (n = 30) was tested on the target and transfer problems to provide a measure of baseline performance.

In the analogy training groups, participants solved letter sequence extrapolation training problems that conformed to the same patterns as those used in the target problems. The single analog group solved one and the same training problem three times and the multiple analog group solved three different training problems once each. Each of the multiple analog problems had different surface features but they all shared the same underlying pattern. In both implicit learning groups, participants memorized letter strings that conformed to the same patterns as those in the target extrapolation problems. The low implicit group memorized six training strings and the high implicit group memorized eighteen training strings. In the control group participants received no training.

General procedure. Participants were tested in groups of 1-4 people. The procedure consisted of *two cycles*. Each cycle was composed of problem training followed by solving target and transfer problems.

Procedure for analogy groups. Participants were first given general instructions on how to solve sequence extrapolation problems. Next, they were presented with the first extrapolation training problem. They were given 6 minutes to solve each problem. After participants had finished solving a problem or max time had elapsed, they were presented with the next problem. After participants solved all three training problems they were presented with the target problem instruction. Target problem instructions were the same as the training instructions except that they added the hint that if participants noticed a pattern on any of the prior problems it would help them solve the target problem. They were then presented with the target problem and were given 6 minutes to solve it. Finally, they were given the transfer problem and instructed to solve it in the same manner as the target problem. The second cycle proceeded in the same manner. The entire procedure took between 60-80 minutes.

Procedure for implicit learning groups. The participants were instructed to memorize and recall each letter string one by one; six strings for the low implicit group and eighteen strings for the high implicit group. They were then given 45 seconds to memorize each string. After 45 seconds the string disappeared and they were given 30 seconds to recall and type in the string. After they finished recalling the string or 30 seconds elapsed, participants were presented with the next string. This procedure was continued through all of the training strings. Next, participants were instructed to write down whether or not they noticed a pattern in the memorization strings. If they noticed a pattern they were instructed to describe it as best they could. Participants were then given general instructions on how to solve the sequence extrapolation problems. They were presented with the target sequence extrapolation problem and given the hint that if they noticed a pattern from the memorization strings it might help them on problem solving. They were given 6 minutes to solve the problem. Finally, they were given the transfer problem and were instructed to solve it in the same manner as the target problem. The second cycle proceeded in the same way. The entire procedure took approximately 70-90 minutes.

Results

The central question of interest is whether or not the various training scenarios facilitated performance on the target and transfer problem tasks. The problem solving score was the number of letters correctly extrapolated in each problem solving task. Because the participants were asked to continue the sequence to eight places their problem solving scores varied between 0 and 8.

Initial analyses revealed non-significant differences within both the implicit and analogy groups and all subsequent analyses collapsed across them, $F(1, 46) = .922$, ns and $F(1, 47) = .02$, ns respectively. Figure 1 shows the mean problem solving scores for the analogy, implicit and control groups on target and transfer problems.

A 3 (treatment: analogy vs. implicit vs. control) by 2 (problem-type: target vs. transfer) mixed analysis of variance (ANOVA) revealed a main effect for both treatment and problem-type, $F(1, 124) = 7.88$, $MSE = 12.96$, $p < .05$ and $F(1, 124) = 14.26$, $MSE = 1.86$, $p < .05$ respectively. The interaction was not significant, $F(1, 124) = .32$, ns. The main effect of problem-type shows that the participants performed better on the target problems than on the transfer problems. Follow up comparisons on treatment showed that the analogy

group performed significantly better than both implicit and control groups, F (1, 95) = 13.77, MSE = 12.35, p < .05 and F (1, 77) = 8.39, MSE = 14.23, p < .05. The implicit group did not significantly differ from the control F (1, 76) = 0.02, ns.

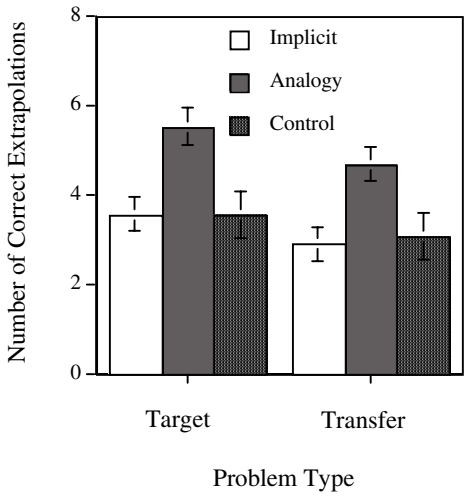


Figure 1: Problem solving as a function of training condition.

In addition, a simple comparison of target vs. transfer was conducted for the control group to provide a measure of baseline performance for problem-type. The analysis revealed a non-significant difference indicating that control participants performed equally well on both target and transfer problems, F (1, 29) = 2.77, ns.

We next compare analogy training to direct instruction.

Experiment 2: Analogy vs. Direct Instruction

Both analogy and implicit learning are indirect training methods. Is it possible to teach a sequential schema directly, by simply telling the participants what the pattern is? Experiment 2 compares different types of analogy training (single vs. multiple analog training) to different levels of direct instruction (high vs. low) on problem solving task performance.

Method

Participants One hundred and nineteen undergraduate students from the University of Illinois at Chicago participated in return for course credit.

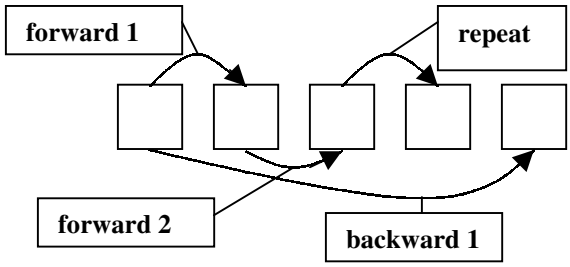
Materials The test problems and the analogy training problems were exactly the same as those used in experiment 1.

In addition, there were two extrapolation problem tutorial booklets, one for low instruction training (12 pages) and one for high instruction training (14 pages). Both tutorials consisted of general instructions for how to *find* pattern sequences as well as detailed descriptions of the component relations of patterns (e.g., forward, backward, repeat and identity relations). The high instruction tutorial had two additional pages of general instruction describing how to *extrapolate* or *continue* sequence patterns. There was also a general tutorial test that consisted of four recall questions and two comprehension questions.

In addition, there were two diagrammatic illustrations of the underlying pattern relations for each of the target problems as well as two blank diagrammatic recall sheets (see Table 3 for an example of a diagrammatic pattern illustration). There were also two distractor tasks that consisted of three multiplication problems each.

Table 3. A sample diagrammatic pattern illustration.

Example: *square boxes represent letter positions*



The test problems and analogy training problems were presented via computer with the same specifications as experiment 1. Direct instruction training material was presented on sheets or in booklet form. Target and transfer problems and associated training stimuli were counter-balanced across all conditions.

Design and procedure The participants were randomly assigned to one of four groups: *single analog* (n = 30), *multiple analog* (n = 31), *low instruction* (n = 28), *high instruction* (n = 30). The same *control condition* (n = 30) was used from experiment 1 as a measure of baseline performance.

In the instruction training conditions participants first read general tutorials, then memorized and recalled the abstract patterns for each target task. The only difference between high and low instruction groups was that the high instruction participants were given two additional pages in the tutorial which provided specific step by step instructions for how to extrapolate a problem.

Procedure for analogy groups. Procedure was exactly the same as in experiment 1.

Procedure for direct instruction groups. Participants were tested in groups of 1-4 people. The procedure consisted of *two cycles*, a training phase and a test phase. Before the training phase cycle all participants were given the general tutorial text to read (max time allowed = 18 minutes) after which they were given the tutorial test (max time = 6 minutes). At the beginning of the training cycle participants were given 3 minutes to memorize the first diagrammatic pattern illustration. Next, participants were presented with

the diagrammatic blank recall sheet and instructed to recall and write down the relations of the pattern (max time = 3 minutes). Participants were then given the distractor task (max time = 3 minutes). Next, participants were presented with the general instructions for the test problems. They were then given 6 minutes to solve the target problem. Finally, they were given the transfer problem and were instructed to solve it in the same manner as the target problem. The second cycle proceeded in the same way. The entire procedure took approximately 70-90 minutes.

Results

Again, the question of interest pursued here is whether or not training facilitated problem solving performance on the test tasks. The problem solving score was calculated in the same manner as experiment 1.

Initial analysis revealed a non-significant difference between analogy groups and all subsequent analyses collapsed across them, F (1, 59) = .51, ns. Figure 2 shows the mean problem solving scores for the analogy, high instruction, low instruction and control groups on target and transfer problems.

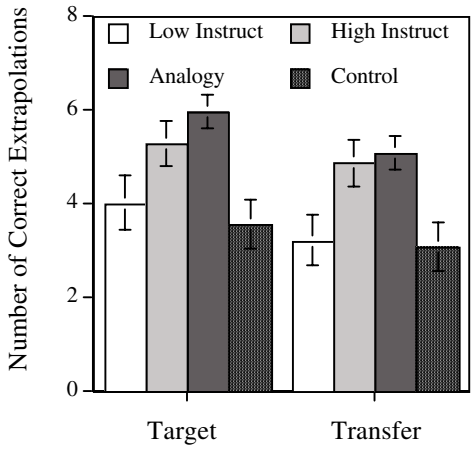


Figure 2: Problem solving as a function of training condition.

A 4 (treatment: analogy vs. high instruction vs. low instruction vs. control) by 2 (problem-type: target vs. transfer) mixed ANOVA revealed a main effect for both treatment and problem-type, F (1, 145) = 6.18, MSE = 14.45, $p < .05$ and F (1, 145) = 18.56, MSE = 1.49, $p < .05$ respectively. The interaction was not significant, F (1, 145) = .70, ns. The main effect of problem-type shows that the participants performed better on the target problems than on the transfer problems. Follow up comparisons on treatment showed that both the analogy and high instruction group performed significantly better than the low instruction and control groups, F (1, 87) = 9.19, MSE = 15.13, $p < .05$, F (1, 89) = 13.11, MSE = 14.77, $p < .05$, and F (1, 56) = 4.37, MSE = 13.94, $p < .05$, F (1, 58) = 6.78, MSE = 13.43, $p < .05$ respectively. The analogy group did not significantly differ from the high instruction group and the low instruction group did not significantly differ from control, F (1, 89) = .60, ns and F (1, 56) = .15, p = ns respectively.

Discussion

So how is abstract, generative knowledge acquired? The present study suggests that there are at least two ways to acquire such knowledge, one through analogical problem solving and the other through direct instruction.

Experiments 1 and 2 showed that participants in the analogy and high instruction training conditions performed better than participants in the implicit, low instruction and control conditions on both target and transfer problems. Target problem performance indicates that the knowledge acquired from analogy and high instruction training was both *generative*, in that the representation could be used to continue a sequence of temporally related actions, and *abstract*, in that the knowledge was flexible and could be applied to novel stimuli. This result supports typical findings on analogical transfer in problem solving (e.g., Gentner & Markman, 1998; Gick & Holyoak, 1983; Reed, 1987).

The analogy and high instruction groups also performed better than implicit, low instruction and control groups on the transfer problems indicating that the knowledge representation was generalizable to similar types of problem structures. However, there was a main effect for problem-type showing that analogy and high instruction participants performed better on target than on transfer problems. In contrast, the control group performed no differently on target than on transfer. These results show that the difference in performance on the target and transfer problem was a function of knowledge gained from training and not of differences in problem stimulus.

These results suggest at least two plausible explanations. One possibility is that some of the participants in the analogy and high instruction groups acquired a knowledge representation of a higher-level abstraction that facilitated their performance on the transfer problem whereas the others did not. Individual differences within the groups would account for the acquisition of a more abstract representation for only a portion of the participants. A second possibility is that participants in addition to learning the abstract pattern from the training stimuli also learned general problem solving heuristics for solving sequence extrapolation problems (i.e., employing specific pattern finding strategies such as searching for repetitions or backward relations). In this case, although knowledge of the specific abstract pattern facilitated performance on target problems it failed to transfer to the transfer problems and participants resorted to more general (and less accurate) problem solving heuristics.

Why did analogy and high instruction training facilitate problem solving and the other training conditions fail? The

prior analysis of the properties of a successful knowledge representation – abstraction and generativity – also reveals the potential components for failure in problem solving including failures of generativity and abstraction.

The failure of implicit training can be explained by either of the above components. Previous research does not provide definitive support for either component. For example, in a prior study we investigated implicit learning in sequence extrapolation problems and found that the knowledge created from the training procedures was of a moderate level of abstraction but was also potentially generative (Nokes & Ohlsson, 2000). Further research is needed to differentiate between each of these components.

The reason for the failure of low instruction can be investigated by examining the differences between the high and low instruction training materials. The only difference between the two training scenarios was that the high instruction tutorial had two additional pages of instruction describing in detail how to *extrapolate* sequence patterns. This description included one example problem that was worked through step by step in detail. This difference in training materials suggests that the low instruction group failed to construct a knowledge representation that was generative.

This hypothesis is also supported by other results in the literature. For example, Ohlsson and Regan (in press) used an intervention paradigm to teach participants several abstract concepts relating to the structure of DNA and facilitated subsequent use of those concepts on a discovery problem. They had participants practice *generating* their own concrete examples of the concepts in the training phase in addition to being given an example from the experimenter. Since low instruction participants in the current study never practiced articulating the abstract schema this component of the knowledge representation was not strengthened.

Thus a commonality that ties both implicit and low instruction together is the lack of practice in *articulating* the abstract schema. Although both learning scenarios had participants memorizing and recalling exemplars, whether they were letter strings or abstract rules, the participants never studied or practiced using these knowledge representations. In contrast, the analogy groups practiced generating pattern sequences on three separate occasions and the high instruction group studied pattern articulation in depth. It is proposed here that it is the practice of pattern articulation of an abstract schema that gives the analogy and high instruction groups their advantage over the other two learning scenarios.

References

Bobrow, D., & Collins, A. (Eds.), (1975). *Representation and understanding*. New York: Academic Press.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy, *Cognitive Science, 7*, 155-170.

Gentner, D., & Markman, A. B. (1998). Structure mapping in analogy and similarity. In P. Thagard (Ed.), *Mind readings* (pp. 127-156). Cambridge, MA: MIT Press

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1-38.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science, 13*, 295-355.

Kintsch, W., & van Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*, 363-394.

Kotovsky, K., & Simon, H. (1973). Empirical tests of a theory of human acquisition of concepts for sequential patterns. *Cognitive Psychology*, *4*, 399-424.

Marshall, S. (1995). *Schemas in problem solving*. Cambridge, UK: Cambridge University Press.

Nokes, T. J., & Ohlsson, S. (2000). An inquiry into the function of implicit knowledge and its role in problem solving. In L. R. Gleitman and A. K. Joshi, (Eds.), *Proceedings of the Twenty Second Annual meeting of the Cognitive Science Society* (pp. 829-834), Mahaw, NJ: Erlbaum.

Ohlsson, S., & Hemmerich, J. (1999). Articulating an explanatory schema: A preliminary model and supporting results. In M. Hahn and S. Stoness, (Eds.), *Proceedings of the Twenty First Annual meeting of the Cognitive Science Society* (pp. 490-495), Mahaw, NJ: Erlbaum.

Ohlsson, S., & Regan, S. (in press). A function for abstract ideas in conceptual learning and discovery. *Cognitive Science Quarterly*.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, *118*, 219-235.

Reed, S. K. (1987). A structure mapping model for word problems. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 13*, 127-139.

Seger, C. A. (1994). Implicit Learning. *Psychological Bulletin, 115*, 163-196.

Simon, H. (1972). Complexity and the representation of patterned sequences of symbols. *Psychological Review*, *79*, 369-382.

Stadler, M., & Frensch, P, (Eds.), (1998). *Handbook of implicit learning*. Thousand Oaks, CA: SAGE.

Thorndike, P. W. (1977). Cognitive structures in comprehension and memory of narrative discourse. *Cognitive Psychology, 9*, 77-110.

Thorndike, P. W. (1984). Applications of schema theory in cognitive research. In J. R. Anderson & S. M. Kosslyn (Eds.), *Tutorials in learning and memory: Essays in honor of Gordon Bower* (pp. 167-192). New York: Freeman & Company.

Thurstone, L. & Thurstone, T. (1941). *Factorial studies in intelligence*. Chicago: University of Chicago Press.

The Age-Complicity Hypothesis: A Cognitive Account of Some Historical Linguistic Data

Marcus O'Toole (marcuso@dai.ed.ac.uk)
Division of Informatics, 80 South Bridge
Edinburgh, EH1 1HN Scotland

Jon Oberlander (jon@cogsci.ed.ac.uk)
Division of Informatics, 2 Buccleuch Place
Edinburgh, EH8 9LW Scotland

Richard Shillcock (rcs@cogsci.ed.ac.uk)
Division of Informatics & Department of Psychology, 2 Buccleuch Place
Edinburgh, EH8 9LW Scotland

Abstract

Shillcock, Kirby, McDonald and Brew demonstrate that there is a significant global relationship between word form and meaning across a substantial part of the lexicon of English. Here, 1705 words were studied to establish how their history in the language related to their participation in the correlation between meaning and form. It was found that the meaning-form correlation was significantly stronger for words with earlier dates of entry into the lexicon, implying that an individual word's meaning-form correlation may develop over time. Changes to individual words may be contingent on the word meanings and word forms in the rest of the lexicon.

Introduction

What is the relationship between a word's form and its meaning? And does age matter to the closeness of a relationship?

This paper addresses the second of these questions by building on previous work which addresses the first. The rest of this section introduces general background to the first question; the next section introduces the specific work upon which we build, and our current hypothesis; subsequent sections outline the methods and results of the current study, and draw a general conclusion.

Kelly (1995) suggested that "the hypothesis that phonological cues are unavailable or that people are not sensitive to them have no 'sound' basis in fact". On the one hand, the interaction between phonological and semantic representations has been widely discussed. On the other, the seemingly intuitive idea that there is a structure-preserving relation between these two aspects of words' representations has been largely ignored.

For instance, Dorffner and Harris (1997) report a model predicting that "although the mapping between word form and meaning is arbitrary… novel pseudo-words will prime concepts corresponding to words that are orthographically similar". They go on to discuss findings that showed that when English speakers are presented with pseudo-words, they tend to have associations with English words similar in terms of form. However, Dorffner and Harris dismiss possible relations between orthographic and semantic representations, and this implies that they see no connection between phonological form and the meaning of words. Yet no strong evidence is put forward to support this claim. Whereas priming effects could certainly exist in the absence of meaning-form relations, there is no reason to suggest that useful phonological cues to meaning cannot be utilised.

Indeed, Kelly (1992) had previously investigated how phonological cues—in terms of number of syllables, word duration, and pronunciation of certain syllables—were involved in category assignment. The study indicated that phonetic cues could be used to infer gender in a number of different languages, including French, Hebrew and Russian.

Continuing from this work, Cassidy, Kelly and Sharoni (*in press*) studied how phonological cues can be used to interpret gender, and how this information might be used by English speakers. A connectionist model was trained to classify novel names as male or female, solely on the basis of phonological cues, and succeeded in classifying 80% of names correctly. Experiments were undertaken which showed that four-year-old children had the ability to infer gender from pseudo-names, and that names that are phonologically typical of either gender are classified significantly more quickly than less typical examples.

In addition, Kelly (1998) studied "blend structure", which concerns the manner in which aspects of two words can be combined to produce a fresh word, in terms of cognitive and linguistic principles. Clearly, such cases help enhance the relationship between meaning and form for the words involved. Blended words such as *brunch* may become embedded in the lexicon due to their phonological evocativeness.

The Relationship Between Meaning and Form

Shillcock, Kirby, McDonald and Brew (2001) report a study in which they generated a semantic hyperspace from a large corpus of English, effectively defining the meaning of each word in terms of its contexts of occurrence. The semantic distance between any two words could be quantified using this hyperspace. In addition, they defined the phonological distance between any two words in terms of an edit distance (the number of features that it would be necessary to change to turn one word in the other).

For a set of 1733 monosyllabic, monomorphemic words, they obtained the meaning distance and the form distance between each word and every other word. They demonstrated that there was, overall, a significant relationship between these two distances: words that are phonologically more similar tend to be semantically more similar.

Further, for each word they calculated the correlation between the two pairs of distances between that word and each of the remaining 1732 words. This gave a value of r_{mf} (the correlation between meaning and form) for each word. When these individual values were ranked, important psychological differences between words emerged between different parts of the ranking. A high value for r_{mf} can be seen as the rest of the lexicon conspiring to support the relation between meaning and form for that particular word. Shillcock et al. claim that the communicatively important words predominantly occurred at the top of the list. In contrast, words with a small or negative r_{mf} value are often more specific and "propositional", for example, *priest* or *plight*. Shillcock et al. postulated that this relation is an example of a tendency towards structure-preserving mappings by the brain.

This ranked list of the relation between words' meaning and form provides the basis for the current study.

With reference to studies on the role of phonology and similarities between lexical neighbours, Shillcock et al. suggested that even very different words can be related in "a model that assumes the whole lexicon may influence the processing of any one word". They go on to demonstrate how the variability of monomorphemic, monosyllabic words in terms of length and phonological similarity can relate to semantic meanings, due to a tendency whose results resemble the compositionality normally present only at the higher levels of linguistic structure.

The Age-Complicity Hypothesis

Suppose the meaning-form correlation is, as hypothesised, a quantifiable aspect of a representational strategy employed by the brain. Then we can suggest that the correlation for each word is open to change over time, as groups of phonetically similar and semantically similar words become established. Further to this effect, it is also likely that words with strong individual meaning-form correlations are more likely to *remain* established, whereas words with weak individual meaning-form relations could be subject to semantic drift, and would be more likely to take on new meanings, losing their original ones. Therefore, the meaning-form correlation would tend to be strongest for words that have been longest in the lexicon. We call this prediction the Age-Complicity Hypothesis.

Method

The current study was based on the list of 1733 words ranked according to their individual relationship between meaning and form (r_{mf}), produced by Shillcock et al. Using information obtained from the Oxford English Dictionary on the first non-obsolete date of entry, a total of 1705 of these words were analysed, once words with no dictionary entry had been discarded. This produced a database of 1705 words with information on first non-obsolete date of entry and r_{mf} scores.

Correlation between r_{mf} rank and date of first entry was measured using Kendall's Tau B. The data were further analysed by comparing how the r_{mf} correlation differs within sections of the ranked list partitioned in terms of r_{mf} rank and age.

Results and Discussion

The Age-Complicity hypothesis predicted that words with long established meanings would have high r_{mf} values. This prediction was convincingly supported by these results. Figure 1 gives an overall view of the data, but although certain interesting features are visible, it is not possible to discern the trends in the data. Figure 2 charts averaged data, and trends become visible.

As Table 1 shows, there was a highly significant correlation ($\tau = 0.08$, $p < .001$) between the meaning-form relation and the first date of entry over all the data.

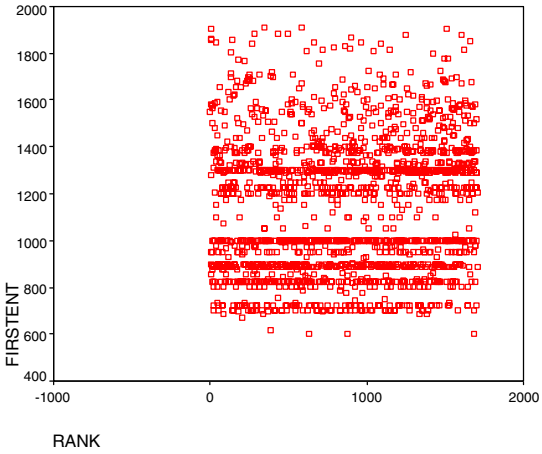


Figure 1: Graph of r_{mf} rank versus date of entry

On the basis of r_{mf} rank, the data were divided into three sections for further analysis: the top ranked 500 words, the middle ranked 500 words and the lowest ranked 500 words. The correlations between first date of entry and r_{mf} are given in Table 1.

There was a negative correlation for the top 500 words in the ranking: the words that were higher in the ranking by r_{mf} had the more recent dates of entry for this subset, contrary to the overall correlation. This feature of the results is probably a reporting phenomenon. The very top of the ranking by r_{mf} contains a number of items such as speech editing terms (*um*, *er*), swear words, and shortened proper names (*Mick*). These kinds of items may have been relatively unlikely to be written down—and hence given a date of first entry—in earlier times.

The middle section showed a non-significant correlation between rank order of r_{mf} value and date of first entry, although this time with a positive correlation. Finally, the lowest entries showed a correlation between rank and first date of entry which mirrored that in the total survey. In summary, despite the anomaly at the top of the ranked list, the results displayed an overall pattern of words' individual meaning-form relationships correlating with age, with words with a high value of r_{mf} having an earlier date of entry.

Table 1: Correlation (τ) between r_{mf} values and date of entry, sorted by r_{mf} rank

Entries 1-1705	Entries 1-500	Entries 600-1099	Entries 1206-1705
0.080 ($p < 0.000$)	-0.056 ($p < 0.033$)	0.032 ($p < 0.148$)	0.080 ($p < 0.004$)

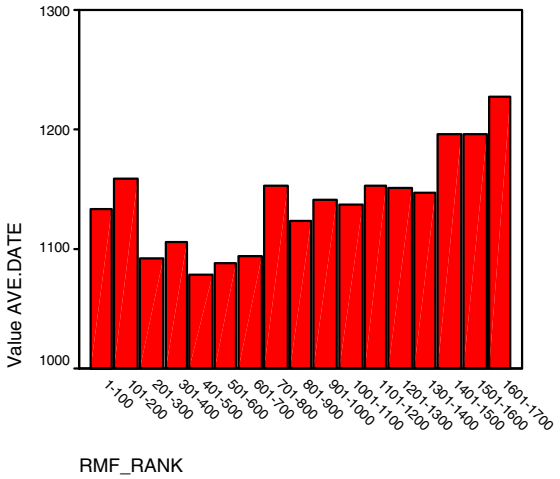


Figure 2: Graph of average first date of entry versus ranked r_{mf} entries

What happens when we think primarily in terms of dates? To provide another view onto the hypothesis, the data were again split into three sections, this time according to dates of entry. The three sections correspond to dates of entry between the years 601 and 897, 972 and 1297, 1303 and 1911, respectively.

Table 2 shows a non-significant (but negative) correlation between r_{mf} rank and date of entry for the section of words with the earliest dates of entry. The positive overall correlation is best reflected in the "middle-aged" words.

We have just seen, in Table 1, the reversal of the correlation for the more recent words, and the same explanation applies to Table 2. A good example is the new entry *yeah* (a colloquial form first recorded in 1905) with a very high r_{mf} value (ranked 8th). If this is due to a reporting phenomenon, then *yeah* could have been important in speech for a long time, and simply not captured in text, to be reported by the OED. On the other hand, it remains possible that *yeah* is just a recent, successful innovation. Then, the entry into the language of such a modified word could be attributed to their meaning-form correlation being stronger than their competitor, (*yes* is ranked 89th in terms of r_{mf}). Figure 3 charts the averaged data.

Table 2: Correlation (τ) between r_{mf} values and date of entry, sorted by date

All Dates 1-1705	Dates ranked 1-476	Dates ranked 599-1100	Dates ranked 1238-1705
0.080 ($p < 0.000$)	-0.043 ($p < 0.091$)	0.064 ($p < 0.023$)	-0.053 ($p < 0.045$)

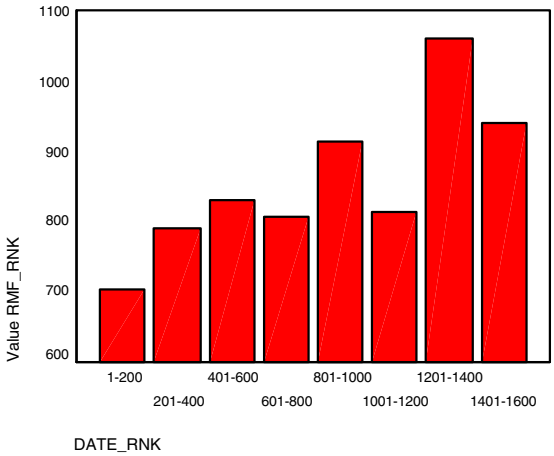


Figure 3: Graph of average r_{mf} rank versus date of entry.

Conclusion

From the results, it seems that there are anomalies amongst the (apparently) youngest, and oldest recorded words in our study. Nonetheless, a moderate interpretation of the Age-Complicity hypothesis has real support: the closeness of the relationship between form and meaning is related to its age.

More specifically, this study has produced evidence of a relationship between the history of an individual word in the language and that word's participation in the overall relationship between meaning and form in the lexicon. If a word has a high value of $r_{mf,}$ then it may be that it resists any change in its own meaning-form relationship; the rest of the lexicon is in effect supporting that relationship. Such a meaning-form relationship is adaptive; it means that the form and the meaning of the word can be partly inferred, one from the other—a clear advantage in language acquisition and processing. At the same time, that individual word may be helping to change the form and/or the meaning of words with weaker values of r_{mf}.

These findings have implications for studies of a number of aspects of human language. Principally, they offer data to substantiate an explanation of why some words become established in the lexicon while others do not. In other words, one of the contributing factors which can help a word become established is "sounding right", i.e. that its form resembles that of words with similar meanings.

We suggest that computational modelling might be used to simulate the data we have presented. It is not possible to obtain sufficient historical data to resolve all the possible reporting biases that may be present in data of the kind we have considered. Computational modelling may help to resolve some of these issues.

Finally, we have shown that it is possible to construe some of the data about language change from historical linguistics in cognitive terms, and specifically in terms of an adaptive relationship between meaning and form in the mental lexicon.

References

Cassidy, K.W., Kelly, M.H., & Sharoni, L. (in press). Inferring gender from name phonology. *Journal of Experimental Psychology: General*

Dorffner, G., & Harris, C. L., (1997) When pseudoword become words – effects of learning on orthographic similarity priming. In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 185-189). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kelly, M.H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review, 99*, 349-364.

Kelly, M.H. (1995). The role of phonology in grammatical category assignments. In J.L. Morgan and K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 249-262). Mahwah, NJ: Lawrence Erlbaum Associates.

Kelly, M.H. (1998). To "brunch" or to "brench": Some aspects of blend structure. *Linguistics, 36*, 579–590.

Shillcock, R., Kirby, S., Mcdonald, S. & Brew, C. (2001). The relationship between form and meaning in the mental lexicon. Unpublished manuscript.

Singular and General Causal Arguments

Uwe Oestermeier (uwe_oestermeier@kmrc.de)
Knowledge Media Research Center, Konrad-Adenauer-Str. 40
D-72072 Tuebingen, Germany

Friedrich W. Hesse (friedrich.hesse@uni-tuebingen.de)
Knowledge Media Research Center, Konrad-Adenauer-Str. 40
D-72072 Tuebingen, Germany

Abstract

In two studies we examined arguments for singular and general causal claims. The first study, a content analysis of newspaper articles, revealed characteristic distributions for mechanistic, statistical, and counterfactual argumentations in singular and general problems. In a second experimental study, subjects formulated arguments from different perspectives on general and singular problems. The results show that subjects are sensitive to the singular-general distinction as well as to related argumentative roles and backgrounds of knowledge. This supports a rhetorical model of causal cognition.

Introduction

Causal arguments consist of causal claims or conclusions and premises which support the causal claims as reasons. The claims may be general (e.g. "smoking causes cancer") or singular (e.g. "John's cancer was caused by his smoking") in nature. Various reasons can be given for such claims: statistical data (e.g. "because smokers take a much higher risk of getting cancer"), counterfactuals (e.g. "because John would not suffer from cancer, if he had not smoked") and mechanistic explanations (e.g. "because the tobacco carcinogen BPDE damages the P53 gene which is critical in the development of lung cancer"). Statistical (Cheng 1993), counterfactual (Lipe, 1991), and mechanistic arguments (Ahn, Kalish, Medin, & Gelman, 1995) have been the mostly discussed justifications for causal claims, but there are other arguments in addition (Oestermeier & Hesse, 2000). Our main question here is: are there systematic relationships between the generality of the claim and the provided arguments?

Several researchers (e.g. Ahn & Kalish, 2000; Hilton, 1995) have stressed the difference between general and singular causal problems, i.e. cases where causal statements are in need of further justification. Unfortunately, the distinction between singular and general problems is not a clear cut one. In the literature on causal cognition this distinction is often illustrated by general phenomena that occur in all relevant contexts alike (e.g. physical laws) on the one hand and particular instances that only occur at specific points in time and place (e.g. a particular car accident) on the other hand. These are, however, only the extremes of a continuum because causal problems range from single events about limited series of occurences to universal general problems. In the following we count only those causal problems as singular where single (or very few events) at a specific point in time and place are involved, and we count all other as general. Several car accidents at different times at a specific junction, for instance, are considered as a general problem albeit a single property of the junction could be the cause of the multiple accidents. Thus our criterion roughly corresponds to the use of singular and plural in the formulation of the causal problem (e.g. "car accident" vs. "car accidents"). The question then is whether this distinction between the singular and general is indeed a crucial one and whether it shows up consistently in varieties of ordinary argumentative discourse. More specifically, we address the following questions: What are the characteristic arguments for general and singular problems? Do subjects expect different arguments from persons with specific argumentative roles and epistemic backgrounds (i.e. general or singular knowledge)?

Previous Research

Answers to these questions directly contribute to the research about causal attribution and argumentation (see Brem & Rips, 2000, for a recent overview). It is an often replicated finding (Kuhn, 1991; Ahn et al. 1995; Slusher & Anderson, 1996) that empirical evidence, i.e. covariation information and statistical data, plays a limited role in the justification of causal claims. Subjects seem to find information about explanatory mechanisms much more useful and convincing. Brem and Rips (2000), however, found that the formulation of empirical evidence increases if the subjects can cite evidence from whatever source they find appropriate. Under such an ideal condition they produce more empirical backings than in those cases were they rely on their own limited knowledge. Brem and Rips (2000) conclude that the neglect of empirical evidence may be largely due to pragmatic restrictions of the availability of appropriate data.

Table 1: A taxonomy of causal arguments

Argument type	Argument schema	Example
Circumstantial evidence	*A caused B, because*	*The poison killed him, because*
1. Spatio-temporal contiguity	B happened at A	a poison bottle was found near the corpse.
2. Co-occurrences	A often occurred together with B	several people taking this poison died.
3. Similarity of cause & effect	A is similar to B	the corpse smells like the poison.
Contrastive evidence		
3. Covariation	B changes with A	the more he took the more ill he became.
4. Statistical covariation	A increases the probability of B	this poison increases the risk to die.
5. Before-after-comparison	B exists after A but not before A	he became ill after taking it.
6. Experimental comparison	action A led to B, action ´A to ´B	poisoned rats died in the experiment.
7. Counterfactual	B would not have happened without A	he would not have died, if he had not taken it.
Causal evidence		
9. Mechanism/causal explanation	A led to C via the process/mechanism B	the poison impaired the metabolism.
10. No alternative	there is no better explanation for B	no other cause leads to such a painful death.
11. Typical effect	B happened and B is a typical effect of A	the colour of the skin is typical for this poison

The cited research remains indeterminate with respect to the relative impact of counterfactual, covariation, and statistical arguments in singular (e.g. a car accident and the death of a plant) and general (e.g. AIDS, unemployment) problems. There are, however, findings that can be taken as a starting point. Kuhn (1991) used general problems and found that subjects formulate mechanistic arguments much more often than covariation arguments, whereas counterfactuals formulations remained nearly non-existent. In singular problems, on the other hand, most subjects also ask for information about mechanism and not for covariation data (Ahn et al. 1995). Lipe (1991), however, found a preference for counterfactual over covariation information and alternative explanations if counterfactual information was available in particular cases.

These findings are compatible with the following hypotheses: Mechanistic arguments occur more often than statistical and counterfactual arguments in general (Kuhn, 1991) and singular cases (Ahn et al., 1995) alike because both kinds of problems are mainly solved on the basis of prior causal knowledge. Counterfactuals are restricted to singular cases because the relevant contrast cases are easier to imagine in concrete cases than in more diffuse global problems (Sherman & McConnell, 1996). Statistical arguments are inherently general and thus restricted to such problems (Ahn & Kalish, 2000, but see Cheng, 1993). With these hypotheses from the literature in mind we looked at newspaper corpora with a great diversity of causal problems.

Study 1: Singular and General Problems

Method

The content analysis of these newspaper articles was based on our taxonomy of causal arguments (see Oestermeier & Hesse, 2000). This taxonomy was developed from several sources: the general philosophical, rhetorical and psychological literature on causation, Kuhn's (1991) interview study, a content analysis of 42 newspaper articles and a pilot study of our own. Table 1 shows the core of our taxonomy, other parts are beyond the scope of this paper.

The first corpus for the taxonomic analysis was taken from the ECI/MCI CD-Rom (the Multilingual Corpus 1 of the European Corpus Initiative). All Frankfurter Rundschau articles were electronically scanned for the keyword "verursach*" (German for "to cause"). This scan of thousands of articles led to a sample of 1024 articles. From this collection a random subsample of 60 articles was drawn. These articles were classified by two independent raters. Cohen's kappa was calculated for the agreement on implicit and explicit causal claims (.81), the segmentation of causal arguments, i.e. whether the text provided complete claim-ground structures (.74), the classification of complete arguments according to our taxonomy (.66), the singular or general nature of the causal problem (.69). The rest of this corpus was analyzed by a single rater.

The method of electronic scanning for a keyword has serious limitations: synonyms, counterfactuals and implicit causal statements are ignored. Besides that all articles of this sample dated from 1992 to 1993. In order to overcome these restrictions two additional printed samples of other newspapers were read by a single rater. One sample of 10 newspaper was randomly drawn from the Schwaebisches Tagblatt (a local newspaper) of the year 1996, another sample of 30 newspapers randomly from the Koelnische Volkszeitung of the year 1903. The rater was instructed to read all articles for all causal arguments with singular and general conclusions and to omit only the parts without journalistic content (like obituary notices, tables etc.)

Results and Discussion

Table 2 shows the frequencies of the argument patterns across singular and general problems. Arguments from explanatory mechanisms were by far the most common ones (75.1%) in both conditions. They were followed by unspecific covariation arguments (4.3%)

and statistical covariations comparing multiple observations (3.5%). All other arguments for causal claims, including counterfactuals, remained below 3%.

72% of the described problems were singular in nature. This gives some support to the hypothesis that singular causal problems are more important for laypeople than general ones (Hilton, 1995). Besides that, singular and general causal problems seem to rely, at least in part, on different reasoning patterns. Explanations of mechanisms dominated singular and general problems alike, but spatio-temporal contiguity (argument 1), experimental comparisons (7) and no alternatives (10) were offered as arguments for causal claims nearly exclusively in single case problems, whereas statistical arguments (5) remained completely restricted to general problems. They occurred only once in the newspaper from 1903. This observation can be best explained by the relatively late penetration of life with statistics which took place after 1900. Argumentations in newspaper certainly reflect historical developments. We would, however, put not too much weight on this hypothesis from one isolated finding.

Some non-findings are also interesting. From the literature we expected that necessary and sufficient conditions should often be used as arguments for causal claims (e.g. Einhorn & Hogarth, 1985). Astonishingly this was not the case, we found no uses of arguments of the form "X caused Y, because all X are followed by Y" or "X caused Y, because Y never happens without X" in the corpora. Even arguments based on counterfactual necessity, although considered as essential for causal reasoning by many researchers (e.g. Lipe, 1991), were rare. The latter may be due to the fact that counterfactuals can be considered as implicit causal arguments that provide at the same time a ground and a claim, whereas our content analysis looked for causal arguments with distinct claims and grounds.

In sum, the data show that singular and general problems cannot be reduced to the same set of argument patterns. Mechanistic or causal explanations that infer causal claims from prior causal knowledge are abundant in both cases, but inductive and abductive arguments are dependent on the problem type. Especially covariation and statistical arguments seem to be restricted mainly to general problems.

Study 2: Roles and Epistemic Backgrounds

The distributions of arguments in the newspaper corpora can only in part be explained by chance and the question is whether subjects are sensitive to these characteristic distributions. Such a sensitivity would be very useful. Argumentation is a complex social activity in which subjects try to defend their interests and gain acceptance by others. The ability to anticipate different arguments from people with different backgrounds of knowledge would offer distinctive advantages in debates. By anticipating certain arguments, for instance, one can prepare the appropriate counter-arguments in advance and thus be in a better position to convince an audience from one's own perspective.

It is clear, however, that subjects can show this rhetorical competence only in settings were it is demanded. Law suits are especially demanding in this respect and therefore we used selected juridical causal problems with various argumentative points of view (plaintiff, defendant) and various backgrounds of knowledge (witnesses, experts).

Hypotheses

We assumed that participants should be able to take the epistemic backgrounds into account and thus expect more often arguments with a reference to concrete spatio-temporal relations from witnesses than from experts.

Table 2. Singular (S) and general (G) causal arguments in three newspaper corpora

Arguments	Volkszeitung (1903)	Rundschau (1992/3)	Tagblatt (1996)	Totals	
	S:G	S:G	S:G	S:G	S+G in %
Circumstantial evidence					
1. Spatio-temporal contiguity	3:0	2:0	5:0	10:0	1.7
2. Co-occurrences	1:1	2:3	0:1	3:5	1.3
3. Similarity of cause and effect		1:0		1:0	0.2
Contrastive evidence					
4. Covariation	5:7	2:4	2:6	9:17	4.3
5. Statistical covariation	0:1	0:10	0:10	0:21	3.5
6. Before-after-comparison	4:2			4:2	1.0
7. Experimental comparison	2:1	8:0	4:0	14:1	2.5
8. Counterfactual	5:4	1:0	1:0	7:4	1.8
Causal evidence					
9. Mechanism	67:26	183:40	94:45	344:111	75.1
10. No alternative	5:0	4:0	1:0	10:0	1.7
11. Typical effect	2:0		2:0	4:0	0.7
Other causal arguments	9:1	21:5	1:4	31:10	6.7
Total S:G	103:43	224:61	110:65	437:169	(100)

In general problems subjects should expect more statistical arguments from experts than from witnesses because it is unlikely that a witness shares the expert's knowledge about statistics. In other words, the participants' arguments should reflect their (perhaps tacit) knowledge that witnesses typically know only about the particular circumstances, whereas experts know about many different cases.

As a generalization of previous findings we expected that mechanistic arguments should dominate singular and general problems alike, because these arguments mirror directly the familiarity of subjects with everyday causal explanations and theories. But singular and general problems should be different with respect to statistical arguments, i.e. arguments that compare multiple observations. We expect more statistical arguments in general than in singular problems, because they abstract from particular circumstances and are inherently general. According to Sherman and McConnell (1996) counterfactual arguments should be occur more often in singular than in general problems.

Method

Participants. The participants were 40 paid volunteer students from various faculties of the University of Tuebingen (16 participants were male and 24 were female; ages varied between 18 and 44 years with a median of 24). Each subject was paid 30 DM. Participants were tested in groups from 3 up to 6 persons and required between 90 and 120 min to complete the paper and pencil tasks. One participant was removed from the data set because of difficulties to understand the instructions and questions in German.

Procedure. We used a mixed repeated measurement design. As a between subject factor 19 subjects had to work on general problems, 20 on corresponding singular ones. The subjects were randomly assigned to these two conditions. As within-subject variables every participant had to work on three different problem contents (food poisoning, allergic reactions, and car accidents) and several perspectives on the case at issue (plaintiff's, defendant's, witness', expert's view and the participant's own pros, cons and final justification). The problem contents and issues were based on real newspaper reports and reformulated for the sake of the experiment. The order of tasks was randomized.

Each participant read three problem descriptions and the related questions. The description of the singular food poisoning case, for instance, read as follows:

> "The organizer of the last year's public festival at the Rhine promenade was sued for compensation for personal suffering at the inferior court Duesseldorf. The plaintiff Oliver K. (36), a visitor of the festival, had suffered from a serious food poisoning. He had taken a snack at one of the snack stands. The vendor could be identified but went into hiding several month ago. An investigation of the case revealed that the vendor had no official license to sell food.
>
> In the last year, the organizer had granted numerous commercial licenses to vendors of snacks and peddlers without checking for the necessary official licenses and health certificates. This was not disputed by the organizer.
>
> At issue between the parties was the cause of the food poisoning. The question of guilt and responsibility was set aside for the moment, at issue was only the question, which cause lay behind this incident."

The first paragraph of the general food poisoning problem read as follows:

> "Several members of the spontaneously founded interest group 'Festival without Fear' brought an action against the municipality of Duesseldorf at the administrative tribunal Duesseldorf in order to lay the city under an obligation to choose another organizer for the traditional yearly public festival at the Rhine promenade. At the last festival numerous visitors had suffered from serious food poisonings. The investigation of these cases revealed that many vendors had no official license to sell food."

The other two paragraphs were identical with the former ones with the exception that plural constructions were used where appropriate.

After the description of the problem scenario the participants were asked for free formulations of causal claims and justifications from different point of views and juridical roles. Single and general case versions were identical except for number and gender:

> "What is the cause of the food poisoning in the plaintiff's [defendant's] point of view? Which justification do you expect from the plaintiff [defendant] for his position?"

In addition, the subject was asked for his own conjecture about the cause and possible pros and cons that would speak for or against his/her conjecture. Each question was followed by three empty lines to allow for free answers in complete sentences. The next sheet started with a causal claim from the plaintiff's view and asked for possible arguments from a witness' and expert's point of view:

> "The plaintiff argues, that the missing controls caused the food poisoning. He cites a witness and an expert.
>
> Which justification(s) do you expect from a witness, who worked at the festival, [an expert, who was procured by the court] for the claim that the missing controls caused the food poisonings?"

Finally, the subject was asked for his/her opinion about the cause and a final argument. The justifications that were freely formulated by the participants were classified in terms of our taxonomy by two independent raters. On a subset of five randomly chosen survey booklets, we calculated Cohen's kappa as a measure of rater agreement. When coding non-causal arguments as a default category (i.e. "other") the result was fair (.54). A second agreement measure was calculated solely for the categories within the taxonomy, i.e. those cases where both rater agreed that the argument in question

was causal in the sense of our taxonomy. This agreement was .76. Non-causal arguments (e.g. arguments from authority) were ignored later on. Differences in classification of arguments were resolved by discussion.

Results and Discussion

Table 3 gives a summary of the results. The overall number of causal arguments which were produced by the participants across all three tasks varied considerably from 5 to 33 with a median of 19.2. To compensate for this, we computed the relative percentages of each produced argument type per person. With the percentages of (a) statistical, (b) counterfactual, and (c) mechanistic arguments as dependent variables we calculated three 2x3x7 (type of task x content x perspective) mixed analyses of variance (ANOVA) with general and singular problems as a between factor and task content and perspective serving as repeated measures.

(a) Statistical arguments were less often formulated (in total 36 formulated arguments) than mechanistic arguments (in total 245) and especially rare under the singular condition (9 out of 36). In consequence there was a main effect of singular vs. general problems for statistical arguments, $F(1, 37) = 4.71$, $p < .05$. There was also a significant main effect of perspective, $F(6, 222) = 3.30$, $p < .01$, i.e. statistical arguments were more often formulated under the perspective of an expert (13) than under all other perspectives (0 for witnesses).

(b) Counterfactual arguments were more often formulated in singular (24) than in general problems (19) but in accordance with our first study and against the hypothesis of Sherman and McConnell (1996) this difference was not significant $F(1, 37) = .747$, $p = .39$. There was, however, a significant interaction between task type and content, $F(2, 74) = 10.894$, $p < .05$. In the car accident and allergic reaction tasks counterfactuals were more often produced under the singular condition than under the general one (14 vs. 5 respectively 5 vs. 1). In the food poisoning problem, this difference was reversed (5 vs. 13). We can offer no explanation for this interaction, but one can argue that our kinds of problems were perhaps not as general as necessary in order to gain a stable effect. Our problems involved multiple people and events but were far from universal because the described events occurred in restricted areas. The interactions show, however, again how content and context dependent causal argumentations are.

(c) Mechanistic arguments were the most common ones (245 in total) and produced by all participants alike across general (130) and singular (115) problems. We found, however, significant main effects of task content, $F(2, 74) = 19.32$, $p < .001$, and perspective, $F(6, 222) = 19.39$, $p < .001$. Nearly half (120) of all mechanistic arguments were formulated in the car accident problem. This may be due to the fact that the participants were more familiar with plausible car accident scenarios than with allergic reactions and food poisonings. In the latter two problem types the mechanisms behind the observable symptoms are less known and hidden to the unaided senses. Against the general trend of a dominance of mechanistic explanations, participants argued from the plaintiff's view equally often with observational before-after-comparisons as with mechanisms (both 20), whereas counter-arguments offering alternative explanations (28 out of 59) were especially frequent under the defendant's view. This shows a clear understanding of the addressed argumentative roles.

Table 3. Frequencies of singular (S) vs. general (G) arguments by perspective and contents

Arguments	**Total**	**Perspectives**					**Contents**		
		Plaintiff	*Defendant*	*Witness*	*Expert*	*Subject**	*Food poisoning*	*Car accident*	*Allergic reactions*
	S:G	S:G	S:G	S:G	S:G	S:G	S:G	S:G	S:G
Circumstantial evidence									
1. Spatio-temporal contiguity	37:25	9:8	4:1	8:5	2:2	14:9	6:3	7:1	24:21
2. Co-occurrences	23:30	0:4	0:2	10:7	4:5	9:12	9:8	1:7	13:15
3. Similarity of cause and effect	0:1	0:0	0:0	0:0	0:0	0:1	0:1	0:0	0:0
Contrastive evidence									
4. Covariation	6:33	1:3	1:3	0:3	1:5	3:19	2:9	2:12	2:12
5. Statistical covariation	9:27	0:5	1:2	0:0	3:10	5:10	5:7	4:13	0:7
6. Before-after-comparison	27:35	11:9	1:0	7:15	2:3	6:8	7:1	1:6	19:28
7. Experimental comparisor	10:7	1:0	1:1	0:1	2:1	6:4	2:1	0:0	8:6
8. Counterfactual	24:19	9:3	3:2	2:4	2:2	8:8	5:13	14:5	5:1
Causal evidence									
9. Mechanism	115:130	7:13	6:11	32:31	33:29	37:46	34:41	56:64	25:25
10. No alternative	6:10	1:2	0:0	1:1	1:3	3:4	4:2	0:4	2:4
11. Typical effect	20:6	3:0	2:1	6:2	6:0	3:3	3:0	4:1	13:5
Other arguments	72:82	6:5	25:21	1:0	2:2	35:33	29:24	16:17	27:18

Note: *This column contains the pros, cons and final arguments of the subject from his/her own perspective

General Discussion

Not all researchers define singular and general arguments in the same way as we do. Basically, we distinguished claims about one and many cases. In the literature a related but different distinction is often drawn in respect to episodic vs. semantic or conceptual knowledge. We decided to use a deviating one-many distinction for several reasons: Firstly, our criterion shows up relatively clearly at the language surface (singular vs. plural) whereas the distinction between episodic and semantic knowledge is much more implicit. Phenomena like unemployment and crime, for instance, can be viewed as local episodic or truly universal problems alike and it often not clear from the context whether causal claims about these problems are intended as propositions about the specific circumstances in a particular economy or society or as general law-like statements. Secondly, frequently cited examples for general claims like "smoking causes cancer" sound to be common but a closer look at non-scientific text corpora shows that such unrestricted causal statements are rare. They seem to be of limited importance in non-scientific contexts.

Thus it remains an open question whether the episodic vs. law-like distinction leads to similar characteristic argument sets. Up to now, relatively few studies looked directly at the verbalization of causal arguments (Brem & Rips, 2000; Kuhn, 1991; Thagard, 1999). A reason for this may be the implicitness and vagueness of ordinary language which makes utterances difficult to analyze. But this obstacle is unavoidable if one wants to understand how causal knowledge is established and communicated in modern societies. Taxonomic and rhetoric studies are indispensable in this respect and in our opinion they should become import guides for further research. The above mentioned studies show, for instance, that statistical (Cheng, 1993) and counterfactual (Lipe, 1991) theories of causal reasoning in psychology have no foundation in a prevalence of the corresponding argument patterns in ordinary discourses. The complex distributions of argument patterns that occur in ordinary language simply cannot be explained from theories that put their focus on a single central and normative causal argument pattern.

Differences in perspectives as well as differences along the singular-general-continuum pose especially difficult problems for reductionistic theories. If the supposed reasoning patterns are the same for all persons, perspectives, and problems alike, it is hard to see how differences can emerge at all. From a rhetorical perspective, however, differences in reasoning are fundamental and in many cases irreconcilable. Our data show that argumentative competencies of humans are highly sensitive to such rhetorical demands, i.e. specific contents and contexts, argumentative roles, different backgrounds of knowledge, restrictions in knowledge, etc. We do not claim, however, that by taking a rhetorical perspective alone, these competencies are already explained. But in our view, it is a progress if such a perspective shift leads to a more adequate description of the pragmatic aspects of causal reasoning and argumentation.

References

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. Cognition, 54, 299-352.

Ahn, W., & Kalish, C. W. (2000). The Role of Mechanism Beliefs in Causal Reasoning. In F. C. Keil & R. A. Wilson (Eds.), Explanation and Cognition. Cambridge, MA: MIT Press.

Brem, S. K. & Rips, L. J. (2000). Explanation and Evidence in Informal Argument, Cognitive Science, Vol 24 (4), 573-604.

Cheng, P. W. (1993). Separating Causal Laws from Casual Facts: Pressing the Limits of Statistical Relevance. In D. L. Medin (Ed.), The Psychology of Learning and Motivation Vol. 30, (pp. 215-264). San Diego: Academic Press.

Einhorn, H. J. & Hogarth, R. M. (1986) Judging Probable Cause. Psychological Bulletin, 99(1), 3-19.

European Corpus Initiative -- Multilingual Corpus 1 [CD-ROM] (1996). Edinburgh: Human Communication Research Centre [Distributor].

Hilton, D. J. (1995). Logic and language in causal explanation. In D. Sperber, D. Premack, & A. J. Premack (Eds.), Causal Cognition: A Multidisciplinary Debate, (pp. 495-529). Oxford: Clarendon Press.

Kuhn. D. (1991). The Skills of Argument. Cambridge: Cambridge University Press.

Lipe, M. G. (1991). Counterfactual Reasoning as a Framework for Attribution Theories. *Psychological Bulletin*, 109(3), 456-471.

Oestermeier, U. & Hesse, F. W. (2000). Verbal and visual causal arguments. Cognition, 75, 65-104.

Sherman, S. J., & McConnell, A. R. (1996). The Role of Counterfactual Thinking in Reasoning. Applied Cognitive Psychology, 10, 113-124.

Thagard, P. (1999). How Scientists Explain Disease. Princeton, NJ: Princeton University Press.

Walton, D. N. (1989). Informal Logic: A Handbook for Critical Argumentation. Cambridge: Cambridge University Press.

Roles of Shared Relations in Induction

Hitoshi Ohnishi (ohnishi@nime.ac.jp)
National Institute of Multimedia Education
2–12 Wakaba Mihama Chiba 261–0014, Japan

Abstract

Two experiments examined the roles of shared relations between representations in induction. Lassaline (1996) found that shared attributes contribute to the inductive strength, but shared relations do not, whereas both shared attributes and shared relations contribute to similarity judgment. A structural alignment view of induction was generalized to account for these phenomena. According to the structural alignment view proposed in this paper, (1) insufficiency of the number of shared relations caused the dissociation between shared relations and inductive strength, and (2) structural alignment during similarity judgment made shared relations so salient as to increase similarity. Experiment 1 examined the first hypothesis. Participants judged inductive strength of arguments that had a crossing number of shared attributes and shared relations. The results showed that shared relations contribute to the inductive strength if a sufficient number of relations are shared. Experiment 2 examined the second hypothesis. The participants who rated similarity between categories of arguments prior to judgment of inductive strength judged arguments having a shared relation to be stronger, whereas the participants who only judged inductive strength did not judge so. The results support the proposed structural alignment view of induction.

Category-based Induction

People frequently make inferences and expand their knowledge in uncertainty. This type of inference is generally referred to as induction. One form of induction where the premises and the conclusion are of the form "All members of a category C have property P" is referred to as category-based induction.

In category-based induction, a categorical argument is said to be strong when the premises increase the degree of belief in the conclusion. Osherson, Smith, Wilkie, Lopez, and Shafir (1990) proposed that the strength of an inductive argument increases with (a) the degree to which the premise categories are similar to the conclusion category, and (b) the degree to which the premise categories are similar to members of the lowest-level category that includes both the premise and the conclusion categories. They implemented their idea as a mathematical model that is called similarity-coverage model. The similarity-coverage model provides a comprehensive explanation to a variety of phenomena in category-based induction.

As an alternative to the similarity-coverage model, Sloman (1993) proposed a connectionist, feature-based, model of induction. According to the feature-based model, an argument whose conclusion claims a relation between category C (e.g., Zebras) and property P (e.g., love onions) is judged strong to the extent that the features of C have already been associated with P in the premises.

Structural Alignment

According to the similarity-coverage model, strength of induction is based on similarities between categories. Recent studies on similarity have revealed flexible and dynamical properties of similarity (Goldstone, 1994; Goldstone, Medin, & Gentner, 1991; Markman & Gentner, 1993; Medin, Goldstone, & Gentner, 1993).

One of the most important findings is that similarity judgment involves a process of alignment of structured relational representations (Markman & Gentner, 1993). In structural alignment, the correspondences between pairs of representations are computed by seeking matches that are structurally consistent. A structurally consistent match means that each attribute or relation in one representation is placed in correspondence with, at most, one attribute or relation in the other representation.

The structural alignment view accounts for important empirical results of similarity and analogy. An important result that is relevant to induction is that attributes and relations in representations are distinguished. This result suggests the possibility that attributes and relations are also distinguished in induction.

Roles of Attributes and Relations in Induction and Similarity

Structural Alignment in Induction

Lassaline (1996) proposed a structural alignment view of induction based on structural alignment of similarity and analogy, and examined roles of attributes and relations in induction and similarity. She hypothesized that (1) shared attributes and shared relation between categories contribute to increasing similarity, whereas (2) nonshared binding relations that connect the target attribute (the attribute being mapped from one category to the other in an inductive judgment) to an attribute shared by the two categories contribute to increasing the strength of inductive arguments.

Inductive arguments related to Hypothesis 1 are illustrated in Figure 1. The four arguments have crossing numbers of shared attributes (2 and 3) and shared relations (0 and 1). In argument (a), two attributes "X and Z" are shared by two animals. Argument (c) is formed by adding a common causal relation to argument (a). Therefore arguments (a) and (c) have two shared attributes, and zero and one shared relation, respectively. Argument (b) and (d) are formed by adding a common attribute to (a) and (c), respectively. Therefore argument (b) and (d) have three shared attributes, and zero and one shared relation, respectively.

According to Hypothesis 1, animal A and B in arguments (c) and (d) are respectively judged more similar than those in arguments (a) and (b) because arguments (c) and (d) have more shared attributes. Similarly, animal A and B in arguments (b) and (d) are respectively judged more similar than those in arguments (a) and (c) because arguments (b) and (d) have a shared relation. Note that shared relations in argument (c) and (d) do not connect the target attribute "Y" and a shared attribute. Therefore these shared relations are not binding relations. Her hypotheses do not make any specific prediction about the contribution of nonbinding shared relations to inductive strength.

(a)
Animal A has X and Z.
Animal B has X, Z and Y.

Animal A also has Y?
(2A-0R)

(b)
Animal A has W, X and Z.
Animal B has W, X, Z and Y.

Animal A also has Y?
(3A-0R)

(c)
Animal A has X and Z.
Animal B has X, Z and Y.
For both animals, X causes Z.

Animal A also has Y?
(2A-1R)

(d)
Animal A has W, X and Z.
Animal B has W, X, Z and Y.
For both animals, X causes Z.

Animal A also has Y?
(3A-1R)

Figure 1: Abstract structure of stimuli used in Lassaline (1996)'s Experiment 1. xA-yR indicates that the number of shared attributes is x, and the number of shared relations is y. W, X, Z, and Y represent attributes of the stimuli.

Hypothesis 2 is relevant to induction. Inductive arguments related Hypothesis 2 are illustrated in Figure 2. The four arguments have crossing numbers of shared attributes (0 and 3) and binding relation (0 and 1). Argument (e) has a shared attribute "X." Argument (g) is formed by adding a nonshared binding causal relation to argument (e). Therefore arguments (e) and (g) have a shared attributes, and zero and one binding relation, respectively. Arguments (f) and (h) are formed by adding two common attributes to (e) and (g), respectively. Therefore argument (f) and (h) have three shared attributes, and zero and one binding relation, respectively.

(e)
Animal A has X and Z.
Animal B has X and Y.

Animal A also has Y?
(1A-0R)

(f)
Animal A has W, X and Z.
Animal B has W, X, Z and Y.

Animal A also has Y?
(3A-0R)

(g)
Animal A has X, W and Z.
Animal B has X and Y.
For animal B, X causes Y.

Animal A also has Y?
(1A-1R)

(h)
Animal A has W, X and Z.
Animal B has W, X, Z and Y.
For animal B, X causes Y.

Animal A also has Y?
(3A-1R)

Figure 2: Abstract structure of stimuli used in Lassaline (1996)'s Experiment 2. xA-yR indicates that the number of shared attributes is x, and the number of binding relations is y. W, X, Z, and Y represent attributes of the stimuli.

According to Hypothesis 2, arguments (g) and (h) are respectively judged as stronger inductive arguments than arguments (e) and (f) because argument (g) and (h) have a binding relation.

She examined roles of attributes and relations in induction and similarity. In her Experiment 1, the roles of shared attributes and shared relations in induction and in similarity were examined. One group of participants rated strength of inductive arguments that had crossing numbers of shared attributes and shared relations as illustrated in Figure 1. The other group of participants rated similarities of pairs of the animals described in each premise of those arguments. The results showed different pattern between inductive strength judgments and similarity judgments. Inductive strength ratings increased by adding a shared attribute, but did not increase by adding a shared relation. In contrast, similarity ratings increased by adding of the shared attribute and the shared relation.

In her Experiment 2, the roles of binding relations were examined. Participants did the same tasks as in Experiment 1 except that arguments included a binding relation as illustrated in Figure 2. The results showed inductive strength ratings increased by adding a binding relation as well as shared attributes. Similarity ratings also increased by adding a binding relation as well as shared attributes.

Both of her hypotheses that were derived from the structural alignment view were supported. Hypothesis 1 is consistent with previous research on similarity(Goldstone, 1994; Goldstone et al., 1991). Hypothesis 2 is consistent with structure mapping theory based on structural alignment of analogy. More specifically, Hypothesis 2 corresponds to systematicity principle in structure mapping theory of analogy (Gentner, 1983).

Roles of Shared Relations in Induction

Lassaline's structural alignment view does not make a specific prediction about roles of shared attributes and shared relations in induction. The results showed that

inductive strength increased by adding a shared attribute. This is consistent with two facts that shared attributes increase similarity and that similarity between categories increases strength of induction.

In contrast, the results about roles of shared relations are problematic. Inductive strength did not increase by adding a shared relation although similarity did increase. Lassaline concluded that a relation must bind the target attribute and a shared attribute to increase inductive strength. However her conclusion cannot explain the results that shared attributes increased inductive strength, but shared relations did not. A hypothesis that binding relations contribute to increasing induction corresponds to systematicity principle in analogy. As most researchers agree, systematicity principle is indeed a strong constraint of analogical mapping, but it is not a unique constraint. In fact, Gentner, Rattermann, and Forbus (1993) revealed that shared relations that are not connected to a common relational structure contribute to soundness ratings of analogy between stories as well as similarity ratings between stories. In addition, shared relations are treated as a constraint of analogy in some computer models (Holyoak & Thagard, 1989; Thagard, Holyoak, Nelson, & Gochfeld, 1990).

In this paper, the structural alignment view of induction is generalized to account for roles of shared relations in induction. Two experiments were conducted to reexamine the roles of shared relations in induction from the viewpoint of a generalized structural alignment. Experiment 1 examined the possibility that the shared relations affect inductive strength. Experiment 2 addressed dissociative roles of shared relations in induction and similarity.

Experiment 1

Goldstone et al. (1991) found that a shared attribute has more weight on similarity than a shared relation when shared attributes are dominant, whereas a shared relation has more weight on similarity than a shared attribute when shared relations are dominant.

The arguments used in Lassaline's Experiment 1 had only zero or one shared relation and two or three shared attributes. Therefore a shared relation might have insufficient saliency on inductive strength. If this hypothesis is correct, shared relations contribute to inductive strength when the arguments have a sufficient number of shared relations.

Method

Participants Fifty-six Keio University undergraduates participated in the experiment as part of the requirements of an introductory psychology course.

Materials and Procedure Each participant was given a booklet that described all tasks and instructions.

Each participant was given a set of 20 pairs of "genotypes of creatures of outer space" that had crossing numbers of shared attributes and shared relations. The number of shared attributes and shared relations were varied from 0 to 4 and from 0 to 3, respectively. Figure 3 shows examples of the pairs. In pair (i) genotypes A and B have a shared attribute and 2 shared relations because both A and B have "◯" in the same place and have the same relations, "on(X, X)": two symbols of the same type are stacked, at the first and third columns. In pair (j) genotypes A and B have a shared attribute and 3 shared relations. The other pairs were constructed in the same manner.

Participants were showed pairs "genotypes of creatures of outer space," referred to as Creature A and Creature B, and were told that each description was intended to refer to a different pair of creatures. They were showed the color of Creature A as a premise. Then they were asked to rate the degree of confirmation that Creature B had the same color as Creature A by selecting a number from 1 (not confirmed) to 9 (completely confirmed) to indicate their judgment.

Four arguments were printed per page. Participants were instructed to spend about 30 seconds in rating the inductive strength of each argument. The orders of the arguments were randomized.

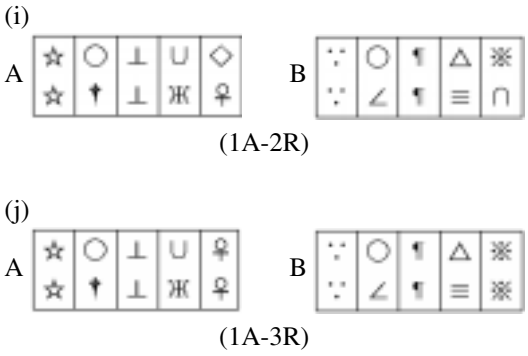


Figure 3: Example of stimuli used in Experiment 1. xA-yR indicates that the number of shared attributes is x, and the number of shared relations is y.

Results and Discussion

Since inductive strength judgments of 7 participants included missing values, they were eliminated from later analysis. As a result, 49 participants' data sets were analyzed.

Results are shown in Figure 4. Inductive strength increased with the addition of shared relations as well as the addition of shared attributes. Mean inductive strength when the arguments had 0, 1, 2, and 3 shared relations were 2.67, 2.95, 3,47, 4.06, respectively. The results showed that inductive strength increased with addition of shared relations. A two-way ANOVA was conducted on inductive strength, with number of shared attributes (0 to 4) and number of shared relations (0 to 3) as within-subject variables.

The ANOVA on inductive strength showed the main effects of number of shared attributes, $F(4,192) =$

$60.12, p <. 01$, and number of shared relations, $F(3,144) = 25.71, p <. 01$.

LSD (Least Significant Difference) post-hoc multiple comparison tests showed that inductive strength increased simply with the addition a shared attribute ($MSe = 2.77, LSD = .33, p <. 05$), and that inductive strength increased simply with the addition a shared relation ($MSe = 3.56, LSD = .34, p <. 05$). As an exceptional case, there was no significant difference between the conditions where number of shared relation were 0 and 1. There was no interaction between number of shared attributes and number of shared relations, $F(12,576) = .578, p > .1$.

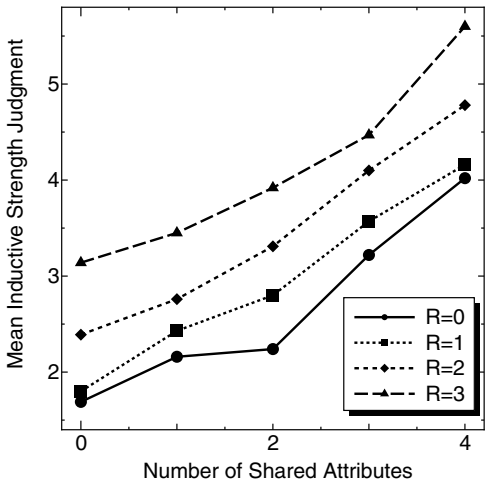


Figure 4: Mean inductive strength judgments from Experiment 1 as a function of number of shared attributes and number of shared relations. $R = 0$, 1, 2, and 3 indicate that numbers of shared relations are 0, 1, 2, and 3, respectively.

The results support the hypothesis that the shared relations contribute to increasing inductive strength judgments when a sufficient number of relations are shared. In addition, there was no significant difference between the conditions where number of shared relation were 0 and 1. This is consistent with Lassaline's results and can be accounted for by insufficiency of the number of shared relations.

A structural alignment view of induction is consistently generalized through Experiment 1. However, the results of Experiment 1 cannot explain dissociative roles of shared relations between induction and similarity because Lassaline's results showed shared relations contributed to increasing similarity but did not contribute to increasing inductive strength. Experiment 2 was conducted to explain these dissociative roles of the shared relations in inductive strength and similarity.

Experiment 2

A structural alignment view of similarity also suggests an explanation for the dissociative roles of shared relations in inductive strength and similarity. Markman and Gentner (1993) proposed that similarity judgment involves a process of structural alignment. A central prediction of structural alignment is that similarity judgments lead people to attend to the matching relational structure in a pair of items. Participants were given a pair of pictures containing cross-mappings where an attribute-based mapping and a relation-based mapping compete, and were asked to select other object in one picture that went with the cross-mapped object in the other. All of these stimuli were explicitly designed so that the participants' natural tendency was to select the similar object that shared an attribute with the other. The participants who rated the similarity of the scenes prior to performing the mapping tasks more often selected the relation-based mappings than the participants who simply performed the mapping tasks without prior similarity judgments.

These results suggest the hypothesis that the shared relations contribute to increasing inductive strength if participants rate similarities prior to inductive strength judgments even when few relations are shared. If the hypothesis is supported, the dissociative roles of shared relations in inductive strength and similarity in Lassaline's results are explained as follows. In inductive strength judgments, since only a relation was shared, a shared relation was not so salient as to contribute to increasing inductive strength. In similarity judgments, a structural alignment during similarity judgment made a shared relation so salient to contribute to increasing similarity.

Experiment 2 examined whether contribution of shared relations to inductive strength increases because of similarity judgments prior to inductive strength judgments.

Participants were assigned to one of the three inductive task conditions. Participants in the "Induction-only" condition performed inductive strength judgment in the same manner as in Experiment 1. Participants in the "Similarity-first" condition first performed similarity rating of categories in the premise and then performed inductive strength judgment. Participants in the "Nonaligning-first" condition first performed the nonaligning task and then performed inductive strength judgment. The Nonaligning-first condition was added in order to rule out the possibility that the difference of inductive strength between Induction-only and Similarity-first conditions was reduced to the difference in the time that participants looked at stimuli.

According to the hypothesis, the inductive strength judged by the participants in the Similarity-first condition is more affected by the shared relation compared with its strength judged by the participants in the two other conditions.

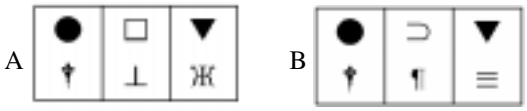
Method

Participants One hundred and sixty Keio University undergraduates participated in the experiment as part of the requirements of an introductory psychology course. They were randomly assigned to one of the three

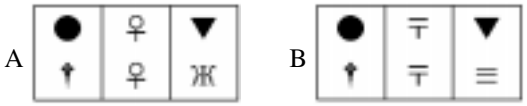
between-subject inductive task conditions.

Materials and Procedure Each participant was given a booklet that described all tasks and instructions in the same manner as in Experiment 1.

Participants in the Induction-only condition performed the same tasks as those in Experiment 1 except for the presented arguments. Each participant was given a set of 8 pairs of "genotypes of creatures of outer space." The number of shared attributes and shared relations were varied from 0 to 3 and 0 to 1, respectively. Figure 5 shows examples of the pairs.



(3A-0R)



(3A-1R)

Figure 5: Example of stimuli used in Experiment 2. xA-yR indicates that the number of shared attributes is x, and the number of shared relations is y.

Participants in the Similarity-first condition first rated similarities of "genotypes" in the arguments by selecting a number 1 (not similar at all) to 9 (very similar) to indicate their judgment. They then performed inductive strength ratings in the same manner as the participants in the Induction-only condition.

Participants in the Nonaligning-first condition first judged whether each creature was an animal or a plant. They then performed inductive strength ratings in the same manner as the participants in the Induction-only condition.

Participants were instructed to spend about 30 seconds in each judgment task.

Results and Discussion

Since judgments of 8 participants included missing values, they were eliminated from later analysis. As result, 51, 51, and 50 participants' data sets in the Induction-only, the Similarity-first, and the Nonaligning-first conditions were analyzed, respectively.

Results are shown in Figure 6. Inductive strength ratings increased by 1.13 points with the addition of a shared relation in the Similarity-first condition whereas inductive strength ratings increased by 0.43 and 0.30 points with the addition of a shared relation in the Induction-only and Nonaligning-first conditions, respectively.

A three-way ANOVA was conducted on inductive strength with inductive task conditions (Induction-only, Similarity-first, Nonaligning-first) as a between-subject

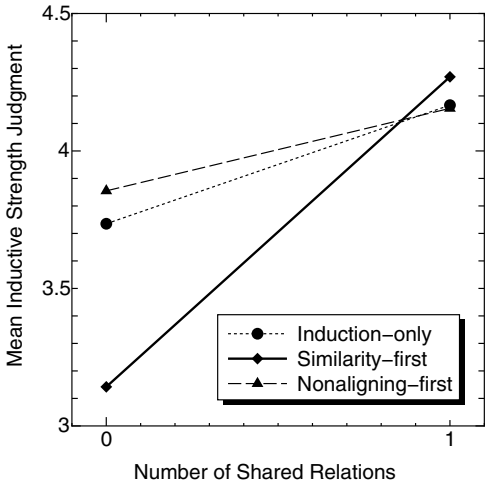


Figure 6: Mean inductive strength judgments from Experiment 2 as a function of number of shared relations.

variable, and with number of shared attributes (0 to 3) and number of shared relation (0, 1) as within-subject variables.

The ANOVA on inductive strength showed the main effects of number of shared relations, $F(1,149) = 38.46, p <. 01$, and number of shared attributes, $F(3,447) = 242.91, p <. 01$. There was no effect of inductive task condition, $F(2,149) = .69, p > .1$.

There was a significant interaction between inductive task conditions and number of shared relation, $F(2,149) = 6.60, p <. 01$. There was also a significant interaction between number of shared attributes and number of shared relation, $F(3,447) = 2.84, p <. 05$. There was no interaction between inductive task conditions and number of shared attributes, nor a three-way interaction ($F(6,447) = 1.49, p > .1$ and $F(6,447) = .51, p > .1$, respectively).

A significant interaction between inductive task condition and number of shared relation supports the hypothesis that similarity judgments prior to inductive strength increase contribution of shared relation to inductive strength. The results cannot be reduced to the difference of the time that participants looked at stimuli because nonaligning tasks did not increased contribution of shared relation to inductive strength judgments.

There were simple main effects of shared relation in the Induction-only and the Nonaligning-first conditions as well as in the Similarity-first condition ($F(1,149) = 6.21, p <. 05$, $F(1,149) = 3.01, p <. 1$, and $F(1,149) = 42.45, p <. 01$, respectively). The most likely explanation for these effects is that pictorial representations of stimuli made participants sensitive to shared relation.

The mean ratings of inductive strength in the Similarity-first condition was lower than two other conditions although there was no effect of inductive task condition. A possible interpretation is that maximal inductive strength was restrained because no pair of stimuli

was not so similar, and the absence of a shared relation decreased inductive strength.

A significant interaction between number of shared attributes and number of shared relation does not have specific interpretation. This interaction was caused by that effects of shared relation were smaller when the shared attribute was zero. The ANOVA was conducted again, this time eliminating the conditions where number of share attributes was zero. The results showed no interaction between number of shared attributes and number of shared relation. The other interactions and main effects did not change.

The results support the hypothesis that the shared relations contribute to increasing inductive strength if participants rate similarities prior to inductive strength judgments even when few shared relations are shared. The results also confirm the explanation for the dissociative roles of shared relations in inductive strength and similarity in Lassaline's results. In her results, a shared relation did not contribute to increasing inductive strength because the shared relation was not salient in this case, whereas a shared relation contributed to increasing similarity because structural alignment during similarity judgment made the shared relation salient.

General Discussion

The structural alignment view proposed by Lassaline (1996) can be consistently generalized as follows. First, attributes and relations are distinguished in induction as well as in similarity judgment. Second, a relation binding the target attribute and shared attributes is a strong constraint in induction. Third, in addition, shared attributes and shared relations are also constraints in induction if they are sufficiently salient. If shared relations are salient, participants easily align relations as well as attributes.

According to the structural alignment view proposed here, shared relations contribute to increasing inductive strength if they are sufficiently salient. In Experiment 1, a sufficient number of shared relations increased to make relations salient. The results showed that shared relations contributed to increasing inductive strength when a sufficient number of relations were shared.

In Experiment 2, participants rated similarity prior to inductive strength judgment to make a shared relation salient. According to the structural alignment view of similarity, similarity judgment involves a structural alignment that leads participants to attend to the matching relational structure. Therefore participants who rated similarity prior to inductive strength judgment were expected to be able to easily align a shared relation. The results showed that a shared relation contributes to increasing inductive strength if participants rated similarity prior to inductive strength judgment.

The results of Experiment 1 and 2 are consistent with the proposed structural alignment view, and are also consistent with the fact that shared relations contribute to soundness ratings of analogy. The proposed structural alignment view of induction more consistently corresponds to a structural alignment view of similarity and analogy than does Lassaline's.

References

Gentner, D. (1983). Structure-mapping: Theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170.

Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, *25*, 524–575.

Goldstone, R. L. (1994). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *23*, 3–28.

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychology*, *23*, 222–262.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, *13*, 295–355.

Lassaline, M. E. (1996). Structural alignment in Induction and Similarity. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *22*, 754–770.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*, 431–467.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*, 254–278.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Reiview*, *97*, 185–200.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, *25*, 231–280.

Thagard, P., Holyoak, K. J., Nelson, G., & Gochfeld, D. (1990). Analog retrieval by constraint satisfaction. *Artificial Intelligence*, *46*, 259–310.

A Model of Embodied Communications with Gestures between Humans and Robots

Tetsuo Ono (tono@mic.atr.co.jp)
ATR Media Integration & Communications Research Laboratories
2-2 Hikaridai, Seikacho, Sorakugun, Kyoto, 619-0288 Japan

Michita Imai (michita@mic.atr.co.jp)
ATR Media Integration & Communications Research Laboratories

Hiroshi Ishiguro (ishiguro@sys.wakayama-u.ac.jp)
Faculty of Systems Engineering, Wakayama University

Abstract

In this paper, we propose a model of embodied communications focusing on body movements. Moreover, we explore the validity of the model through psychological experiments on human-human and human-robot communications involving giving/receiving route directions. The proposed model emphasizes that, in order to achieve smooth communications, it is important for a *relationship* to emerge from a mutually entrained gesture and for a *joint viewpoint* to be obtained by this relationship. The experiments investigated the correlations between body movements and utterance understanding in order to confirm the importance of the two points described above. We use robots so that we can control parameters in experiments and discuss the issues related to the interaction between humans and artifacts. Results supported the validity of the proposed model: in the case of human-human communications, subjects could communicate smoothly when the relationship emerged from the mutually entrained gesture and the joint viewpoint was obtained; in the case of human-robot communications, subjects could understand the robot's utterances under the same conditions but not when the robot's gestures were restricted.

Introduction

Why do people use gestures when communicating? A common scene on a street involving giving/receiving route directions is some person and a stranger making gestures together as if dancing synchronously and rhythmically (see Figure 2). These gestures appear not only when the person describes turns at visible locations, but also at invisible ones. Moreover, it has been shown that people are unable to achieve smooth communications if they are restricted from using spontaneous gestures (Ono et al., 2001). Consequently, gestures play an important role in human-human communications.

In this paper, however, we do not discuss *emblem gestures* such as the OK sign; these gestures have arbitrarily defined meanings and figures in social conventions. The target of our research is *mutually entrained gestures*, where a speaker and a hearer spontaneously and synchronously move their bodies according to the entrainment resulting from mutual actions and utterances. We focus on such gestures because smooth communications between humans can be expected when these gestures are used, as illustrated by the above example involving giving/receiving route directions.

In order to investigate the mechanism of the mutually entrained gestures described above, we conduct experiments on human-robot communications as well as human-human communications. The reason why we use a robot is that we can unrestrictedly design experiments by using a programmable robot's gestures. Moreover, an investigation of human-robot communications can contribute to research on the methodology of robot design and the interaction between humans and artifacts.

The purpose of this paper is to propose a model of embodied communications that can give an explanation for the mechanism of communications described above and, moreover, provide evidence for the validity of the model through psychological experiments. The main characteristic of our model is to focus on the *relationship* emerging from a mutually entrained gesture and the *joint viewpoint* obtained by the relationship. The experiments concretely investigate the correlations between body movements and utterance understanding in human-human and human-robot communications involving giving/receiving route directions. In such a task, it is hard for a person and a stranger to communicate with each other if they do not share the same viewpoint. Here, in order to obtain a joint viewpoint, both sides need to construct a relationship emerging from mutually entrained body movements. We investigate the process of communications in the experiments by using our implemented robot.

Embodied Communications

Previous Research on Gestures

Research on gestures conducted to investigate the mechanism of communications emerged around 1980. In this field, McNeil was the first to carry out cutting-edge research. He pointed out that gestures are synchronized with speech in communications, and thus both are closely connected in the cognitive system (McNeil, 1987). McNeil's research provided findings leading to the development of research on the functions of gestures.

However, previous research has mainly analyzed the speaker's gestures in communications. In other words, many researchers have analytically investigated the correlations between speech and the speaker's body movements. Consequently, their aim has been to explain an internal mechanism of an individual speaker. However,

these research works have not looked into the dynamical mutual interaction between a speaker and a hearer.

In contrast, we focus on the *dynamical mutual interaction* in human-human and human-robot communications involving giving/receiving route directions. Especially, the reason why we come to adopt this route directions is that spontaneous gestures such as pointing easily appear in this context. Kita (2000) analyzed a speaker's gestures for this task but did not deal with the dynamical mutual interaction between them. In this research, we are able to give evidence for a hypothesis in detail because we can control the parameters in experiments by using a robot.

Model of Embodied Communications

In this paper, we propose a model of embodied communications focusing on entrained body movements. Our model is basically described by the following formula:

$$(S, U) \quad I$$

Here, is a viewpoint for understanding an utterance in a situation, S is the situation around a speaker and a hearer, U is an utterance from the speaker, and I is information obtained by having understood utterance U.

For example, let us suppose that in a situation S involving giving/receiving route directions, a person A utters "Go right" to a stranger B while both are facing each other. Let us further suppose that B understands from his/her viewpoint of A that the utterance means the "right" of A. In this case, the relation among the viewpoint, situation, utterance, and information is expressed as $_A(S, U_A) \quad I_B$. However, B may instead understand from his/her viewpoint of himself/herself that the utterance means the "right" of him/her. In this case, the relation is expressed as $_B(S, U_A) \quad I_B$.

The above ambiguity can be effectively solved by using an absolute coordinate system. For example, a person can clearly direct a stranger to a destination when both sides can use a visible landmark or object, or when both sides can construct a similar cognitive map (this assumes the stranger has previously visited the area). In this case, the viewpoint is determined definitely.

However, a person cannot use an absolute coordinate system when landmarks and turns to the destination are invisible, or when a stranger has not visited the area before. In this case, it is difficult to maintain a joint viewpoint in communications because the stranger is unable to imagine the route map of the person; the person's memory access also becomes overloaded.

As described in the Introduction, people seem to solve the problem of deciding a viewpoint by mutually entrained body movements. In other words, people first construct a relationship that emerges from a mutually entrained gesture. This relationship allows people to obtain a joint viewpoint. Finally, they can communicate with each other smoothly because of the utterance understanding achieved as a result of this joint viewpoint.

In our proposed model, the characteristic of communications discussed above is expressed as follows:

$$(\emptyset, U) \quad I \tag{1}$$

$$O({}_iR_j) \tag{2}$$

$$E(torso, arms, eyes) \quad {}_iR_j \tag{3}$$

Here, $\emptyset$ indicates the situation where there is nothing to point out, ${}_iR_j$ is the relationship between persons i and j, and O is a function for obtaining the viewpoint from the relationship. Moreover, *torso* and *arms* are expressions for entrained movements of the torso and arms, while *eyes* expresses the eye contact in communications. E is a function of the relationship emerging from the entrained movements.

These formulae express the process of communications involving giving/receiving route directions as follows. People cannot adopt an absolute coordinate system when they do not have a landmark or object to point out. Consequently, it is hard for them to achieve utterance understanding because of the difficulty of obtaining a joint viewpoint (Formula 1). To overcome this problem, they try to construct a relationship to obtain the joint viewpoint (Formula 2). This relationship emerges from mutual entrained body movements (Formula 3). Smooth communications can be achieved through these processes because the joint viewpoint makes both the speaker's utterance and the hearer's understanding easier.

In our model, we formalize the process of communications described above. We carry out psychological experiments to explore the validity of the model in the following two chapters.

Human-Human Communications

Experiments

Experiments on human-human communications were conducted by the following method.

Outline of experiments We focused on the interaction between a subject and a person involving giving/receiving route directions as an informant just happened to be passing by. Here, we investigated the appearance of their gestures, gestural arrangements, utterances, and the level of utterance understanding.

Subjects Ten undergraduate and graduate students (male and female). The subjects had not previously visited the experimental environment, and thus did not know the route to any destination at all.

Environment Figure 1 shows an outline of the experimental setup. These experiments were done in the hallways of a laboratory. Point A denotes the place where the route directions were given, and B and C denote the goals, i.e., a cafeteria and an information desk, respectively. Point T1 denotes a turn in the route from A to B, and Points T2-T4 denote turns from A to C. Only the corner of T1 is visible from A.

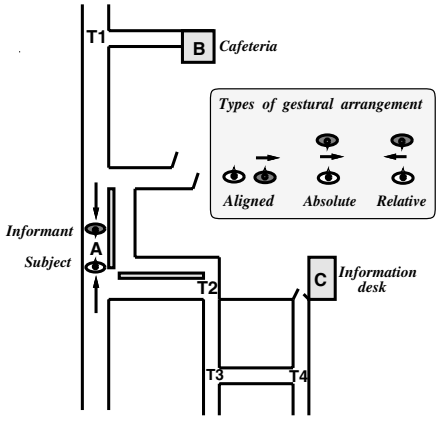


Figure 1: Experimental setup: arrangement of subject, informant, destinations, and turns.

Procedure The subjects received the following instructions from the experimenter (position A): "Ask a passerby the way to the cafeteria and the information desk and go to each place by yourself." The behaviors and utterances of the subjects and the persons were recorded with a camera and a microphone.

Evaluation The results of the experiments were evaluated from the record of the subjects' and the persons' behaviors, i.e., gestural arrangements, arm and elbow movements, and eye contact. In addition, we evaluated the time needed to communicate and the accuracy of the communicated information.

We classified the gestural expressions and arrangements into three categories, i.e., *aligned*, *absolute*, and *relative* gestures, following the literature (Kita, 2000) (see the upper right-hand side of Figure 1). To illustrate, let us assume that, at position A in Figure 1, an informant directs a subject to destination B by telling him/her to turn right at corner T1. In *aligned gesture*, an informant makes gestures to indicate his/her right aligning his/her torso orientation with that of the subject. In *absolute gesture*, an informant makes gestures to indicate the subject's right while facing the subject. In *relative gesture*, an informant makes gestures to indicate his/her right while facing the subject.

Results

First, the gestural expressions and arrangements that the subjects and the persons took were *aligned gesture* in nine out of ten cases and *relative gesture* in the remaining one case. These results were the same for both destinations. Next, Table 1 shows the analyzed results of synchronized gestures between the subjects and the persons. Synchronized arm gestures were observed in six out of ten cases in the route directions to destination B and eight out of ten cases to destination C. Here, synchronized arm gestures mean that the subject synchronously makes similar movements to the person's spontaneous arm movements (see Figure 2). In this experiment, all of the persons made arm movements. In addition, all of those who made synchronized gestures moved their extended arm right and left. Moreover, in all cases, they made eye contact. In particular, in the more complicated route to C, eye contact was made with high frequency.

Table 1: Results of entrained actions of arms and eyes in human-human interaction.

	Arm synchronized	Elbow extended	Eye contact (total)
Cafeteria	6/ 10	6/ 6	12 times/ 123 sec
Information	8/ 10	8/ 8	25 times/ 216 sec

Figure 2: Photo of mutually entrained gesture in human-human communications in the route direction.

Furthermore, the time needed to communicate was 17.2 seconds in the case of destination B but 32.2 seconds in the case of destination C. That is, the more complicated route direction statistically needed much more time ($t_{(18)} = 2.122,\ p < .05$). However, there was not much difference in the kinds of expressions used in the utterances between the two destinations. Eventually, all of the subjects could arrive at the two destinations. In other words, information was accurately communicated from the person to the subjects.

A summary of the experimental results is as follows. First, the persons acting as informants made spontaneous gestures not only when they described turns at visible locations but also at invisible ones. The subjects involuntarily made entrained and synchronized gestures to the persons.

We can assume the following relation between the experimental results and our proposed model. The subjects had not previously visited the experimental environment. Therefore, it was hard for the subjects to understand the persons' utterances because of the difficulty of obtaining a joint viewpoint (Formula 1). To overcome this problem, they tried to construct a relationship to obtain the joint viewpoint (Formula 2). This relationship emerged from mutually entrained body movements (Formula 3).

In the next chapter, we describe experiments on human-robot communications in order to investigate these mechanisms in detail. We can unrestrictedly design the experiments by using a programmable robot's gestures. Moreover, the investigation of human-robot communications can contribute to research on robot design and the interaction between humans and artifacts.

Human-Robot Communications

Experiments

Experiments on human-robot communications were conducted by the following method.

Outline of experiments We focused on the interaction between a subject and a robot as an informant involving giving/receiving route directions. Here, we investigated the appearance of the subject's gesture and the level of utterance understanding while changing the robot's gesture.

Subjects Thirty undergraduate and graduate students (male and female). The subjects were randomly divided into six groups. The subjects had not previously visited this experimental environment, as in the human-human experiments.

Robot Our robot system can make gestures by using the upper part of its body in the same way as a human (see Figure 3). The robot has two arms, two eyes, a mobile platform, and various actuators and sensors. With this equipment, the robot can generate almost all of the behaviors needed for communications with humans.

Environment Figure 4 shows an outline of the experimental setup. These experiments were done in the hallways and lobby of a laboratory. Points S and R denote the initial positions of the subject and robot, respectively. Point A denotes the place where the route directions were given, and B denotes the goal, i.e., the lobby. Points T1-T4 denote turns in the route from A to B, directed by the robot. Only the corner of T1 is visible from A.

Procedure The experiments consisted of the following six phases.

1. The subjects received the following instructions from the experimenter (position S): "Ask the robot the way to a lobby and go there by yourself." The question to the robot was specified as follows: "Tell me the way to the lobby."

2. The subjects moved from S to A, and the robot from R to A.

3. At position A, the subjects asked the question, and the robot answered. The robot could make its utterance with synthesized speech sounds. The content of the utterance was "Go forward, turn right, turn left, turn right, turn left, and then you'll be at the destination." The robot could make gestures while uttering this. In these experiments, we prepared six conditions under which the robot's gesture was changed.

4. The subjects tried to go to the lobby after receiving the robot's directions.

5. The experiments finished whether the subjects arrived at the lobby or gave up after losing their way. The subjects psychologically evaluated the robot through a questionnaire after the experiments finished.

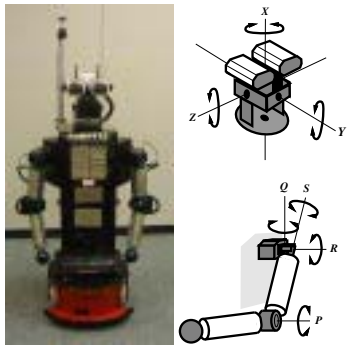


Figure 3: Outline of robot called "Robovie" (left), and robot's head and arm motion mechanisms (right).

Conditions We prepared the following six conditions from C-1 to C-6, which differed in terms of the robot's body movements (see Figure 5). The content of the utterance was the same under every condition.

C-1 (No gesture): The robot did not move.

C-2 (Absolute gesture): The robot raised its left arm leftward when telling the subject that he/she should go right, while it raised its right arm rightward when telling the subject that he/she should go left.

C-3 (Absolute gesture with gaze): In addition to C-2, the robot turned its eyes to the subject while making the utterance.

C-4 (Only aligned torso): The robot rotated so that it aligned its torso with the subject.

C-5 (Aligned gesture): In addition to C-4, the robot raised an arm forward telling the subject when he/she should go forward, rightward when the subject should go rightward, and leftward when the subject should go leftward.

C-6 (Aligned gesture with gaze): In addition to C-5, the robot turned its eyes to the subject while making the utterance.

Evaluations The results of the experiments were evaluated from the record of the subjects' behaviors and the answers of the questionnaire. In the questionnaire, the subjects were asked whether they understood the robot's utterance and to give a psychological evaluation of the robot on a seven-point scale for six items: *Familiarity*, *Sincerity*, *Reliability*, *Intelligence*, *Extroversion*, and *Kindness*.

Predictions

In the experiments, we gave evidence for the following three predictions derived from the proposed model. The more the robot's gestures increase systematically rather than randomly, i.e., the more the conditions shift in order from C-1 to C-6,

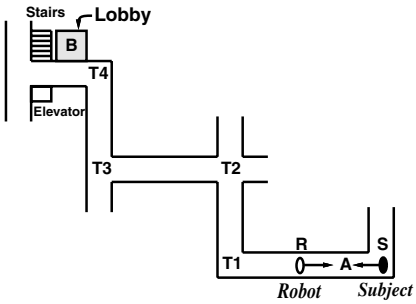


Figure 4: Experimental setup: arrangement of subject, robot, destination, and turns.

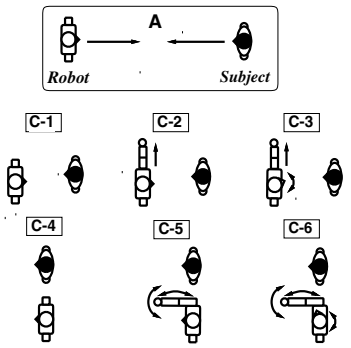


Figure 5: Outline of experimental conditions under changing robot gestures.

Prediction 1: the more the subjects' gestures will increase by entrainment and synchronization with the robot's, and consequently, a relationship will emerge from the mutual gestures.

Prediction 2: the easier the joint viewpoint will be obtained by the relationship.

Prediction 3: the easier the subjects will understand the utterance of the robot and arrive at the destination by using the obtained viewpoint.

Here, Predictions 1, 2, and 3 correspond to Formulae (3), (2), and (1) in the model of embodied communications, respectively.

Results

We give evidence for the three predictions in the order of Predictions 1, 3, and 2 to make the point of our argument clearer.

Verification of Prediction 1 From the observation results on the subjects' behaviors, we analyzed the subjects' gestures. First, the gestural arrangements that the subjects took were as we had expected (see Figure 5). Next, Figure 6 shows the ratio of appearances of the subjects' body movements under each condition. In this analysis, we classified the subjects into three categories: subjects who did not practice body movements at all (*Nothing*), subjects who only moved their hands (*Hand*), and subjects who moved and raised their hands

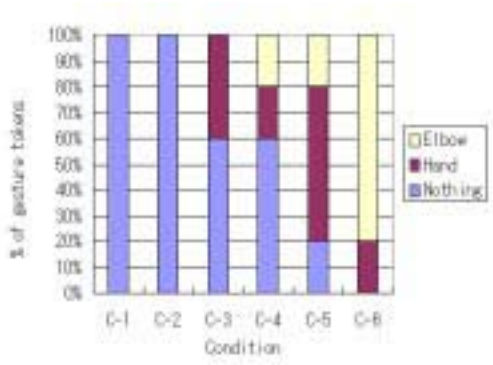


Figure 6: Results of subjects' body movements in human-robot interaction.

Figure 7: Photos of a subject under Condition C-1 (left) and two subjects under C-6 (center and right).

up to the elbow level (*Elbow*). In the analysis, a significant difference was found between the ratio of appearances of the subjects' body movements and the conditions ($\chi^2 = 25.210$, $p < 01$). In other words, the more the conditions shifted from 1 to 6 (i.e., the more the robot's gestures increased systematically), the more the subjects' gestures increased in sync. Moreover, the average scores for the numbers of times the subjects turned their eyes to the robot were higher when the robot turned to meet the eyes of the subjects (C-3 and C6).

We show appearances of the experiments in Figure 7. First, the left-hand side of Figure 7 shows the appearance of a subject not making any body movement and not turning his eyes to the robot at all (C-1). In contrast, the center of Figure 7 shows the appearance of a subject making an entrained body movement and turning his eyes to the robot (C-6). The right-hand side of Figure 7 also shows the appearance of a subject making an entrained body movement and turning her eyes in the same direction as the robot (C-6).

As a result of the observations described above, we could confirm that relationships emerged between the subjects and the robot from mutually entrained gestures. Consequently, Prediction 1 was supported.

Verification of Prediction 3 We recorded the time the subjects spent moving from A to B in Figure 4. Table 2 shows the average time and the number of subjects not arriving at B under each condition. Regarding the average time, no significant difference was found between the conditions. However, the average time in C-6 was the shortest.

Table 2: Average time until subjects' arrival at destination, and number of subjects not arriving at destination.

	C-1	C-2	C-3	C-4	C-5	C-6
Time to destination	69.5	71.3	67.7	70.2	66.8	65.4
Number of subjects not arriving	1	2	2	0	0	0

A noteworthy point is that a considerable number of subjects did not arrive at the goal in C-1, C-2, and C-3. The results of the questionnaire clearly showed that the subjects who did not arrive were unable to correctly understand the robot's utterance. One of the comments often heard was that they could not understand whether the robot's utterance including "left" and "right" meant the robot's or the subjects'. In other words, the reason why the subjects did not understand the utterance was that they could not obtain a joint viewpoint with the robot.

Consequently, the subjects who did manage to obtain a joint viewpoint could understand the robot's utterance and arrive at the goal, whereas the subjects who did not were unable to understand and arrive at the goal. Therefore, Prediction 3 was supported.

Verification of Prediction 2 First, we discuss obtaining a joint viewpoint from the aspect of body arrangement. Under the verification of Prediction 3, it was clear that all of the subjects could obtain a joint viewpoint when the robot aligned its body arrangement with the subject's to the destination (C-4, C-5, and C-6). In contrast, approximately one-third of the subjects could not obtain a joint viewpoint when the robot did not align its body arrangement (C-1, C-2, and C-3). Consequently, it is hard for subjects to obtain a joint viewpoint when no relationship emerges from the use of body arrangement.

Next, we discuss obtaining a joint viewpoint from the aspect of mutually entrained gestures such as synchronized arm movements and eye contact. As discussed in the verification of Prediction 1, from the results of the observed data, the more the robot's gestures increased systematically, the more the subjects' gestures did so. Moreover, from the results of the questionnaire, the more the conditions shifted from C-1 to C-6, the higher the average scores became (see Figure 8). In other words, the more the conditions shifted, the smoother the communications became. Based on this consideration, the relationship that emerged from the entrained gesture made it easier to obtain the joint viewpoint.

As a result of the above observations, Prediction 2 was supported. Consequently, the validity of our proposed model was given evidence by the three supported predictions.

Discussion and Conclusions

In this paper, we proposed a model of embodied communications focusing on body movements. Moreover, we explored the validity of the model through experiments on human-human communications involving giving/receiving route directions. The results of the experiments roughly supported our research direction. However, we could not investigate the details of the model because we were unable to manipulate the parameters in the experiments. Therefore, we carried out similar experiments using our implemented robot system. From the results of these experiments, we could give evidence for the validity of the model more appropriately.

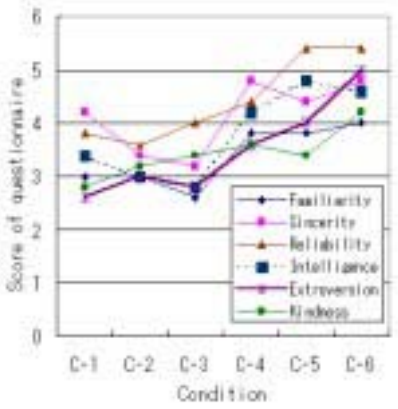


Figure 8: Results of subjects' psychological evaluations to a robot.

The contributions of our research should be viewed from two perspectives. First, our model of embodied communications suggests a new direction in research on communications. The target of previous research had mainly been the mechanism of verbal communications based on informatics approaches, e.g., Shannon's model. After that, McNeil's school pointed out that gestures are synchronized with speech. However, they have not yet modeled a whole conception of interactive communications that includes the function of embodiment. Our model gives a clue toward better understanding of such communications.

Moreover, the results of this research can be applied to interactive technologies between humans and artifacts. In other words, artifacts that can draw out human physical movements can make humans feel familiar with them. These cognitive engineering technologies enable us to develop an interface system and a robot system in the work's next generation.

References

Kita, S. (2000). Interplay of gaze, hand, torso orientation and language in pointing. In *Pointing: Where language, culture, and cognition meet*. Cambridge University Press, Cambridge.

McNeill, D. (1987). *Psycholinguistics: A new approach*. Harper & Row.

Ono, T., Imai, M., Ishiguro, H., & Nakatsu, R. (2001). Embodied communication emergent from mutual physical expression between humans and robots. *Transactions of Information Processing Society of Japan*, (submitted).

Remembering to forget: Modeling inhibitory and competitive mechanisms in human memory.

Mike W Oram (mwo@st-andrews.ac.uk)
Malcolm D. MacLeod (mdm@st-andrews.ac.uk)
School of Psychology, University of St. Andrews, St. Andrews
Fife, KY16 9JU, UK

Abstract

Given the importance attached to memory in everyday life, the inability to recall items on demand can be problematic. An apparently ironic phenomenon has been identified, however, which suggests that in addition to retrieving desired memories, the act of remembering inhibits or suppresses related memories. We show here that a competitive model, designed to investigate the development of the cortical visual system, provides an explanation for the suppression of some memories as a consequence of remembering others. We confirm a number of specific predictions based on our model as to when retrieval-induced forgetting effects should or should not occur. The model suggests that the mechanisms by which memories are formed and adapted may also underlie retrieval-induced forgetting effects. In addition to having important practical implications, the model provides a theoretical base for the transfer of theories and ideas between two separate levels (cortical processing and memory formation and adaptation) of understanding brain function.

Introduction

Recent evidence suggests that far from being a detrimental process, forgetting has an adaptive role (Anderson & McCulloch 1999; Bjork 1989; Macrae & MacLeod 1999). When trying to remember a specific memory, available retrieval cues are often insufficiently specified to the extent that related but unwanted information is also accessed. This unwanted information can interfere with our ability to retrieve the information we wish to recall. A potential solution to this problem is through the temporary suppression or inhibition of related material (Anderson & McCulloch 1999; Anderson & Spellman 1995; Anderson et al 1994; Anderson, Bjork & Bjork 2000; Bjork et al 1998; Ciranni & Shimamura 1999; MacLeod & Macrae 2001; Macrae & MacLeod 1999). Importantly, this temporary suppression of related memories – retrieval-induced forgetting – occurs without the need for explicit cues to forget and can therefore be considered an intrinsic part of the act of remembering (Anderson & Spellman 1995; Anderson et al 1994; Macrae & MacLeod 1999). Other explanations, such as output interference (where items recalled early in a list can interfere with the retrieval of subsequent items) have been eliminated as potential explanations for this phenomenon using a variety of methods. Direct evaluation using statistical techniques have shown that there is no tendency for the retrieval-induced forgetting effect to be larger for those participants who recalled practised items first (MacLeod in press; MacLeod & Macrae 2001; Macrae & MacLeod 1999). More direct evidence that an inhibitory process is involved comes from the demonstration that temporary suppression is observed in all items that are related (whether by initial set or other semantic links) to the suppressed items (i.e. second order inhibition, Anderson & Spellman 1995).

Retrieval-induced Forgetting

In an experiment showing retrieval-induced forgetting, participants are typically given two sets of information to remember regarding two separate categories (*A_1, A_2,…,A_10, B_i, B_ii, B_iii,…,B_x, e.g, 'John_cheerful, John_tolerant,…; Bill_vigorous, Bill_sensible,…)* followed by a retrieval practice session on a subset of items from one of the lists (the retrieval practice or RP set, *A_1, A_2,…A_5,* e.g. complete the following: *'John_ch____').* Following a distracter task (name as many capital cities as you can), participants are asked to recall as many of the items as possible.

Figure 1 shows the pattern of results from such an experiment (see Methods). A greater proportion of the practised items (RP+, left bar) were recalled than unpractised items in either the same set (RP-, middle bar) or in the unpractised set (NRP, right bar). This enhancement (RP+ versus NRP) shows the facilitatory effect of practice on subsequent recall. Retrieval-induced forgetting is evidenced by the fact that recall performance of the non-practised items in the practised set (RP-) was worse than the recall of non-practised items in the non-practised set (NRP). Thus, retrieval-induced forgetting is a selective suppression of related items and not a general suppression of all memories (Anderson & McCulloch 1999; Anderson & Spellman 1995; Anderson et al 1994, 2001; Bjork et al 1998; Ciranni & Shimamura 1999; MacLeod & Macrae 2001; Macrae & MacLeod 1999). An output interference explanation would predict that retrieval-induced forgetting effects would be higher where there was a tendency to recall RP+ items early in the list. As noted in the introduction, retrieval-induced forgetting is not due to items from the practiced subset (RP+ items)

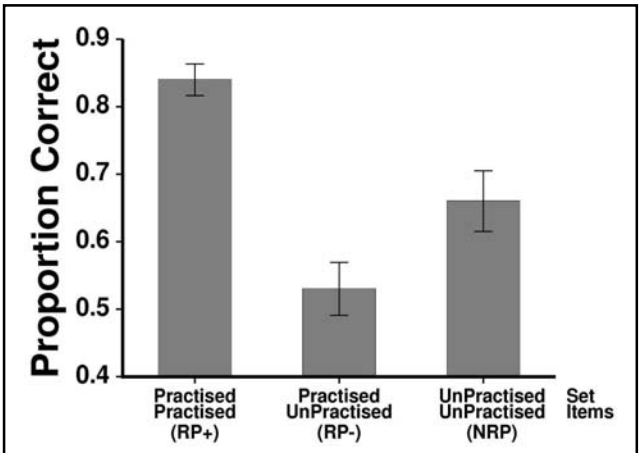


Figure 1: Retrieval-induced forgetting. Mean (±SEM, n=20) of the proportion of items remembered in each of each item type (RP+, RP- or NRP). The recall of unpractised items in the practised set (RP-) was less ($p < 0.05$) than the recall of the unpractised items in the unpractised set (NRP). Overall effect of conditions $F_{[2,38]}=28.3$, $p < 0.0005$.

being recalled first during the free recall task. (Anderson & Spellman 1995; MacLeod in press; MacLeod & Macrae 2001; Macrae & MacLeod 1999).

As the act of remembering during practice can selectively suppress memories for related but unpractised items, retrieval-induced forgetting must be influenced by the relationships between the items established during memory formation (Anderson & McCulloch 1999; Anderson & Spellman 1995). We were therefore interested in whether or not the mechanisms underlying retrieval-induced forgetting were different from the mechanisms that established the memories. We begin by considering, in broad terms, the required properties of a model consistent with the experimental data on retrieval-induced forgetting. As retrieval-induced forgetting is undirected (Anderson & McCulloch 1999; Anderson & Spellman 1995; Anderson et al 1994; Macrae & MacLeod 1999), learning should be unsupervised. Also, as retrieval-induced forgetting occurs with both semantic (Macrae & MacLeod 1999) and episodic memories (Ciranni & Shimamura 1999), the model should show unsupervised learning of both semantic and episodic-like memories. Finally, as inhibitory mechanisms are implicated, the model should contain inhibitory or competitive processes. We first show that a model consistent with this broad outline shows retrieval-induced forgetting. We then use the model to formulate three predictions of when retrieval-induced forgetting will not be observed. These predictions are verified experimentally.

Methods

We employed a computational approach to aid understanding of the role of inhibitory mechanisms in mental life. Computational testing of psychological theories can provide a powerful conceptual framework from which principled sets of research questions can be derived. However, using computational models in this way is not straightforward. The high number of degrees of freedom can lead to over-fitting the data and hence offer neither explanatory power nor generalisation to other scenarios. Hence, the observation that a model can fit experimental data is insufficient to validate the underlying processes within the model. In addition, results from a model developed around underlying psychological processes will be restricted in interpretation to the assumed underlying psychological processes: such a model can determine whether the assumed processes could underlie observed phenomena, but is weak at determining whether the assumed processes are actually in operation and important.

We address the caveats of using computational models to investigate psychological processing in two ways. First, selection of the category of model is made in broad terms without specific implementation to match observed psychological phenomena. If such a model is observed to produce the phenomena of interest, predictions from changing parameters in the model can then be validated with experimental data. The experimental validation of predictions overcomes, at least partially, the difficulties associated with many degrees of freedom that, in turn, gives rise to over-fitting the experimental data. Second, we assume that if the model reflects the psychological processes in a meaningful way, the parameters of the model will relate to psychological processes. This is not simply that the output of the model relates to the phenomena of interest, but that the parameters relate to underlying psychological processes. If the parameters of a model can be related to psychological processes, then the model may provide insight into how these processes interact.

Simulation methods

Damage to cortical tissue appears necessary for retrograde amnesia, implying that the neural representation in cortex correlates with the long-term memory. As inhibitory mechanisms are implicated in both the formation and functioning of neural representation (Oram and Perrett, 1994; Desimone and Duncan, 1995) and cognitive interactions between those representations (Anderson et al., 1994; Anderson and Spellman, 1995; Ciranni and Shimamura, 1999), we chose to investigate whether inhibitory processes involved in the formation of representations/memories could also underlie the interactions between representations/memories revealed by retrieval-induced forgetting.

The model consisted of two sets of 10 input nodes representing the individual items and two input nodes representing the set identifiers. The 22 input nodes were fully connected to the 10 output memory nodes, initially with random weights (0..1). Each node had an associated trace activity, Tr, which was dependent on the node's Tr at the previous time step and the node's current activity, Act: $Tr_{(time+1)} = (1-\delta)Tr_{time}+\delta Act$. The trace activity time constant δ was set at 0.5, with similar results obtained for δ=0.2 to δ=0.8. Weights between input i and memory node j were set randomly (0…1) with updating (learning) based on the trace activity, $\Delta Wt_{[i,j]} = \alpha(Act_i - Wt_{[i,j]})Trace_j$. The weight change rate, α, was 0.01 (similar results were obtained for α=0.001 to α=0.2). The $(Act_i - Wt_{[i,j]})$ ensures that the weights are bounded (-1..1). Initial training consisted of setting the activities of the input node corresponding to one of the input items to 1, calculating the activity of the memory nodes, updating the weights, then resetting the activity of the input node to 0, then "presenting" another input item. The activity of the set node associated with each input item was set to 1 while items within the set were presented. Retrieval practice was run in an analogous way for one half of the items in set 1, except that activity of the item nodes was set at 0.5 representing the partial cueing in the experimental paradigm. The representational strength was calculated as the activity in the item nodes following activation of 1.0 of the set node. Weight change was calculated as the change in the representational strength from after training to after retrieval practice. The change was normalised by dividing by the representational strength after training.

Experimental methods

Following Anderson et al (1994), the study comprised four phases: study, practice, distracter and final test. Participants were presented with ten items of information presented individually for 5s about two witness statements (one concerning a personal theft and the other a bank robbery). The practice phase followed immediately after the study phase. Participants were presented with five questions about a subset of items concerning one of the witness statements (RP+ items). Each question was presented three times. Counterbalancing and randomisation of question order ensured that each of the items appeared equally often in the RP+, RP-, and NRP conditions. Participants were then given a 5-min distracter task (write down as many capital cities as you can). Finally, participants were given a surprise free recall task in which they were required to recall as much of the information contained in the two statements. The number of correctly recalled items was noted for each group (RP+, RP- and NRP) and converted to proportion correct by dividing by the number of items in each group (RP+=5, RP-=5, NRP=10 and Figure 1).

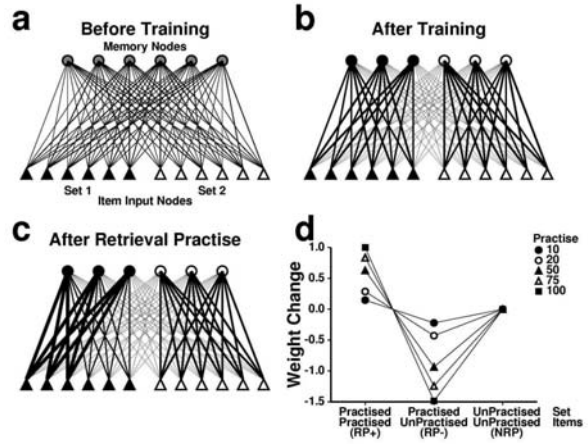


Figure 2: Modelling retrieval-induced forgetting. The weights between item nodes and memory nodes (representing the strength of the memory) are depicted by the thickness of the connecting line. The grey scale of the memory nodes indicates the strength of connection between the set node (not shown) and the memory nodes. A: Before training, the weights between input and memory nodes were random. B: Weights after the training. C: Retrieval practice re-distributed the weights linking the set 1 items with the memory nodes while leaving the weights on set 2 unaltered. D: Change in the weights representing the memories of the items and their set membership. As practice increases from 10 to 100 iterations, the strength of the memory inhibition also increases.

Results

We adapted a fully connected single layer unsupervised competitive model that forms both semantic and episodic like memories by learning from both past and present activity (Foldiak 1990, 1991; Oram & Foldiak 1996). The model consists of two sets of input items, each containing 10 items. Two additional inputs were used to indicate the training set. Initial weights from the input to output nodes were set randomly. Competitive interactions were modelled using a winner-take-all implementation (Foldiak 1991; Oram & Foldiak 1996) between the 10 nodes in the output layer. The network we adapted loads each output node with equal share of the input variance (Foldiak 1990; Oram & Foldiak 1996). The trace activity (Tr) imparts a structure to the inputs in the form of temporal co-variance between items. This co-variance results in the equally distributed input variance being parceled into equal variances associated with the different input sets and, within each set, an equal representation of the individual items. Thus, each output node learns part of the co-variation between a "set" node and the "item nodes". The resulting representation is best described as sparse, being neither fully distributed nor local. Sparse

representations have the benefits of both distributed and local representations and seem to describe accurately cortical representations. The greater the number of output nodes, the sparser the representation. Qualitatively similar results are obtained when the number of output nodes varies from 4-30 output nodes.

There were two phases to training: in the 1st phase, the model was sequentially presented with each of the items with the items set membership also activated. This is analogous to the initial learning phase of retrieval-induced forgetting paradigms. In the 2nd phase, the model is sequentially presented with half the items from one set partially activated (the retrieval practice phase). The changes in the strength of the model's representations of items at different stages during simulated retrieval-induced forgetting are shown schematically in Figure 2a-d (thick lines indicate a strong link, thin lines indicate a weak link). Before training (Figure 2a) the weights are random and small. Learning rules based on recent as well as current activity, such as those employed here, learn temporal relationships between inputs (episodic-like memories) as well as relationships between nodes with concurrent activity (semantic-like memories). This allows the individual set-item relationships and the relationships between the different items within the same set to be learned. The inhibitory competition between nodes keeps the set-item representations of different sets of inputs separate (Foldiak 1991; Oram & Foldiak 1996). After training (Figure 2b) the representation of the

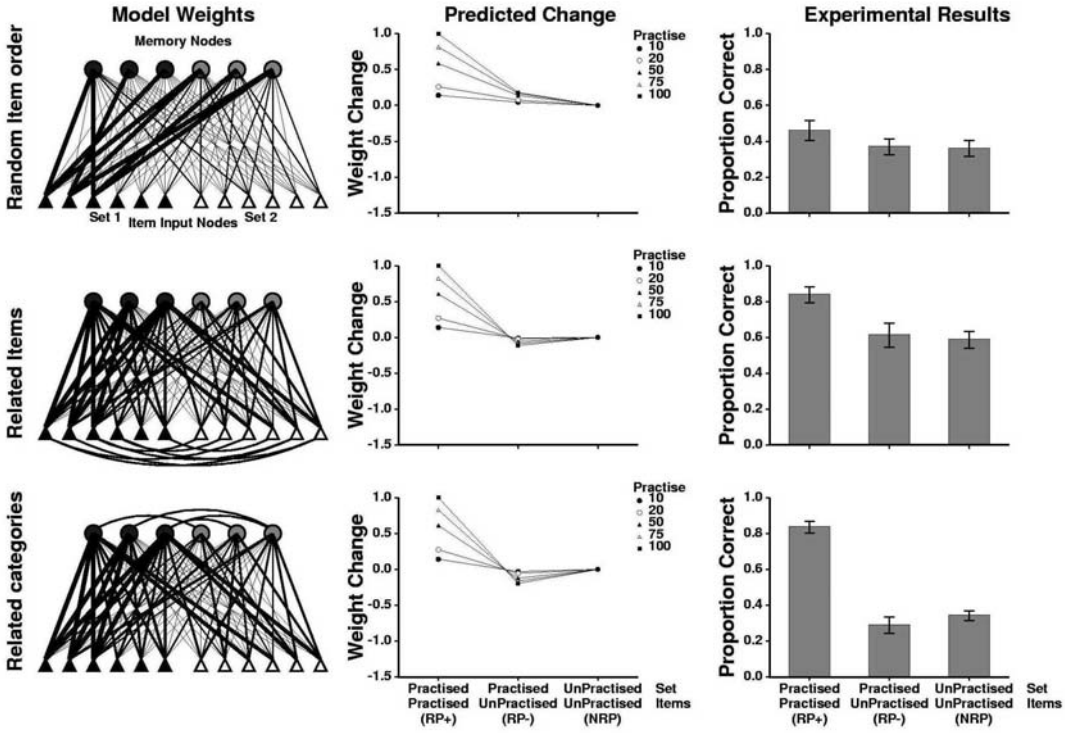


Figure 3: Predicting the disruption of retrieval-induced forgetting. The three rows show different conditions under which the model predicts that retrieval-induced forgetting will not occur and the results of experimental studies. The change in the representational strengths of the RP+, RP- and NRP items in the model following retrieval practice are shown in the middle column. The results of the experimental studies (mean proportion correct ±SEM in free recall for the RP+, RP- and NRP item types) are shown in the right column. Upper Row: Lack of coherence between items (random presentation of items). The model was run with trace activity time constant δ=1 (otherwise experiment as in Figure 1) producing overlapping representations of set 1 and set 2 (left). Accuracy of predicted recall was reduced compared with items which were temporally coherent. The absence of retrieval-induced forgetting (compare top middle with Figure 2D)was confirmed experimentally ($p > 0.5$, right). Middle Row: Effect of semantic links between items. Direct connections (strength = 0.5) between input items in set 1 and set 2 produces overlapping representations of set 1 and set 2 (left). Retrieval-induced forgetting was so attenuated that it was predicted to be undetectable experimentally (middle). This was confirmed experimentally ($p > 0.2$). Lower Row: Effect of confounding by category. When connections between memory nodes (25% chance) were included (strength = 0.5), item-set memories showed overlap between sets (left), which again predicts greatly attenuated effects of retrieval-induced forgetting (middle). This was also confirmed experimentally ($p > 0.05$).

items in the memory nodes is divided into two sets, with little or no overlap. Following retrieval-practice (Figure 2c), the strength of the representation of practised items (RP+) is increased (without the simulated retrieval practice, the representation of the RP- and RP+ items is equivalent). The strength of the representation of unpractised RP- items is reduced because of repeated occurrences of high activity in the memory nodes with no activity in the input nodes representing the RP- items. As the retrieval-practice did not activate the memory nodes associated with the NRP items, the strength of representation items in the non-practised set is not influenced by retrieval-practice. Figure 2d shows that the network predicts the phenomena associated with retrieval-induced forgetting: the expected recall of RP+ items is enhanced compared to the recall of the NRP items and, as with retrieval-induced forgetting, the recall of the RP- items is lower than the recall of the NRP items. Thus, a competitive model can show retrieval-induced forgetting effects. The model does not provide a direct prediction of any effect of the order in which items will be recalled: we use the strength of representation as our metric. Note, however, that although the representational strength would suggest free recall beginning with RP+ items, the retrieval-induced forgetting effect in the model is not due to any form of output interference, only the strength of representation.

Given the variable success experimenters have had in producing retrieval-induced forgetting effects, we were particularly interested in examining the conditions under which retrieval-induced forgetting would not occur. We wished therefore to evaluate the model by changing those model parameters which suggest a strong influence on retrieval-induced forgetting and experimentally testing the predicted effects. We chose to manipulate those model parameters that have readily identifiable psychological counterparts. If the model parameter that represents the degree of continuity between the items (trace activity) is reduced, the items in each of the different sets to be remembered do not form a coherent pattern after the initial training, and the clarity of the relationships between items and their sets is reduced (Figure 3, top left). The overlap between the representations of items in different sets predicts reduced levels of recall performance compared with the situation where the continuity between items is easily established. Following simulated retrieval-practice, the strength of both the RP- and the NRP item representations is reduced while the strength of the RP+ representations is increased, i.e., an absence of retrieval-induced forgetting (Figure 3, top middle). When items used in the initial experiment (Figure 1) were presented in random order such that no coherence between the items was evident, retrieval-induced forgetting did not occur. In addition, absolute performance levels were reduced compared to when the same items were presented in a coherent fashion (compare Figure 3, top right and Figure 1).

Links between individual input items of the different sets can be thought of as exemplar-exemplar links based on semantic relationships between item inputs. Activation of one item will lead to (partial) activation of those related items in the second set. The concurrent activation leads to item representations that do not map perfectly with the input set (Figure 3, center left), so that retrieval practice reduces the strength of representation of both the RP- and NRP items whilst increasing the RP+ representation (Figure 3, center). Semantic relationships between input items were obtained experimentally by using appearance descriptors concerning two individuals (e.g. *Bill_Nike trainers, Bill_Slim build…, John_Adidas trainers, John_Medium build*) as the input items (*trainers*, *build* etc forming explicit links). As predicted, retrieval-induced forgetting did not occur (Figure 3, center right). Finally, links between the representations of the item groups (the memory nodes) models the existence of pre-existing groupings involving the items. This can be thought of as the existence of indirect or implicit semantic links (exemplar-category-exemplar). The overlap of pre-existing groupings of the items of the different sets leads to the representation of single items being associated with both sets (Figure 3, lower left). The effect of confounding relationships between the memory nodes was examined by asking participants to learn representations of employees in different companies that were confounded by gender. Again, the prediction from the model was met: retrieval-induced forgetting did not occur (Figure 3, lower right).

Discussion

The results of these studies highlight two important aspects of memory formation and maintenance. First, we have shown a mechanism by which practice and revision (consolidation) of selected memories can lead to suppression of related memories but leave unrelated and unpractised memories unaffected (Figure 1). While others have noted the restricted occurrence of retrieval-induced forgetting (Anderson & McCulloch 1999), our model allows specific predictions to be made about both performance levels and the strength of retrieval-induced forgetting effects. The four predictions about performance in a cognitive memory task (Figure 3, middle column) were all tested and verified experimentally (Figure 3, right column). This suggests competitive models with learning based on past as well as present activity can help predict how, why and when these types of memory interactions occur. Second, the model suggests that the effects of practice and revision of selected memories are due to the same processes by which memories are first established and hence need not be regarded as separate cognitive processes. Support for retrieval-induced forgetting as an intrinsic

property of memory formation comes not simply from the demonstration that a model can produce retrieval-induced forgetting effects without explicitly coding the effect, but also that the same model predicts the absence of retrieval-induced forgetting effects (Figure 3).

In day-to-day function, retrieval-induced forgetting is important because it allows the updating or alteration of memory without interference of or disruption to other memories. For example, remembering where you parked your car today rather than where you had parked it yesterday should not interfere with your memory of the shopping you need to do. This type of selective adjustment of memories has practical implications: police interview techniques could be adjusted to minimise the potential loss of pertinent information from witnesses; teaching the establishment of conceptual links between aspects of the curriculum should be emphasised with revision of all the related material; students who revise only part of their course may well be placing themselves at a disadvantage because of the active suppression of related memories. If, as our model suggests, retrieval-induced forgetting effects are intrinsic to memory formation, then a simple way of reducing susceptibility to this kind of forgetting is to create many links during initial learning – perhaps the reason why the development of complex schemata provides resistance to such forgetting (Anderson & McCulloch 1999).

We have shown that a competitive model reveals a potential mechanism allowing prediction of experimental data concerning the cognitive processes of memory formation and adaptation. Our model shares similarities with that of Bauml (1997). However, our model suggest the suppression normally attributed to retrieval processes could itself be part of the mechanism by which memories interact and are updated. The choice of model provides not only a potential explanation of memory formation and adaptation but also demonstrates that a mechanism proposed to describe the selectivity of single cells within extra-striate visual cortex (Foldiak 1991; Oram & Foldiak 1996; Oram & Perrett 1996; Wallis & Rolls 1997) can operate at the much coarser scale associated with episodic and semantic memories and their interactions. The existence of a single model that operates at both fine (single cell) and coarse (episodic and semantic memory) scales is appealing because it provides a medium for the transfer of theories and ideas between two different levels of approach to brain function and their subsequent testing.

Acknowledgments

We acknowledge Dr. J. Gerry Quinn for helpful discussion and reading of draft manuscripts.

References

Anderson, M.C., Bjork, R.A. & Bjork, E.L (1994) Remembering can cause forgetting: Retrieval dynamics in long-term-memory. *J Expl Psychol-Learn Mem Cogn* 20, 1063-1087.

Anderson, M.C., Bjork, R.A. & Bjork, E.L. (2000) Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychon Bull Rev*, 522-530.

Anderson, M.C. & Neely, J.H. (1996) in *Memory: Handbook of perception and cognition* (eds Bjork, E.L. & Bjork, R.A.) 237-313 (Academic Press, New York.

Anderson, M.C. & McCulloch, K.C. (1999) Integration as a general boundary condition on retrieval- induced forgetting. *J Exp Psychol Learn Mem Cogn* 25, 608-629.

Anderson, M.C. & Spellman, B.A. (1995) On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. *Psychol Rev* 102, 68-100.

Bauml, K.H. (1997) The list-strength effect: Strength-dependent competition or suppression? *Psychonom Bull Rev* 4, 260-264

Bjork, R.A. (1989) in *Varieties of memory and consciousness: Essay in honor of Endel Tulving* (eds Roediger, H.L. & Craik, F.I.) (Erlbaum, Hillsdale, NJ).

Bjork, E.L., Bjork, R.A. & Anderson, M.C. (1998) Varieties of goal-directed forgetting. In (eds J.M. Golding & C.M. MacLeod) *Intentional forgetting.* Mahway: NJ Erlbaum.

Ciranni, M.A. & Shimamura, A.P. (1999) Retrieval-induced forgetting in episodic memory. *J Exp Psychol Learn Mem Cogn* 25, 1403-1414.

Foldiak, P. (1990) Forming sparse representations by local anti-Hebbian learning. *Biol Cybern* 64, 165-170.

Foldiak, P. (1991) Learning invariance from transformation sequences. *Neural Comput* 3, 194-200.

MacLeod, M.D. (in press) Retrieval-induced forgetting in eyewitness memory: Forgetting as a consequence of remembering. *Applied Cognitive Psychology* In press.

MacLeod, M.D. & Macrae, C.N. (2001) Gone but not forgotten: The transient nature of retrieval-induced forgetting. *Psychol Sci* 12, 148-152.

Macrae, C.N. & MacLeod, M.D. (1999) On recollections lost: When practice makes imperfect. *J Pers Soc Psychol* 77, 463-473.

Oram, M.W. & Foldiak, P. (1996) Learning generalisation and localisation: Competition for stimulus type and receptive field. *Neurocomputing* 11, 297-321.

Oram, M.W. & Perrett, D.I. (1994) Modeling visual recognition from neurobiological contraints. *Neural Networks* 7, 945-972.

Wallis, G. & Rolls, E.T. (1997) Invariant face and object recognition in the visual system. *Prog Neurobiol* 51, 167-194.

The Origins Of Syllable Systems : An Operational Model

Pierre-yves Oudeyer
Sony Computer Science Lab, Paris, France
e-mail : py@csl.sony.fr

Abstract

Many models, computational or not, exist that describe the acquisition of speech: they all rely on the pre-existence of some sort of linguistic structure in the input, i.e. speech itself. Very few address the question of how this coherence and structure appeared. We try here to give a solution concerning syllable systems. We propose an operational model that shows how a society of robotic agents, endowed with a set of non-linguistically specific motor, perceptual, cognitive and social constraints (some of them are obstacles whereas others are opportunities), can collectively build a coherent and structured syllable system from scratch. As opposed to many existing abstract models of the origins of language, as few shortcuts as possible were taken in the way the constraints are implemented. The structural properties of the produced sound systems are extensively studied under the light of phonetics and phonology and more broadly language theory. The model brings more plausibility in favor of theories of language that defend the idea that there needs no innate linguistic specific abilities to explain observed regularities in world languages.

Introduction

There are many studies about the acquisition of speech sounds, and of language in general: a lot of data is available and a lot of theories as well as operational models have been developed (Altman, 1995). Although there is great disagreement for these questions, one assumption is logically done by these models: a pre-existing language already exists, with all its associated structure and redundancies. Depending on the theoretical position, either the acquisition of a particular language consists in adjusting a number of parameters of an innate language acquisition device that already knows most of the structure of languages (Chomsky and Halle, 1968), or it relies on statistical learning techniques able to infer regularities from the data. On the contrary, very little is known about how this structure and these regularities originated, from a situation where there was no sound system at all (and no language in general) ?. In brief, how did speech emerge and why does it have the shape it has ? This ignorance is due partly because the question of the origins of language has been an actively researched question only for slightly more than a decade (Hurford et al. 1998), partly because no meaningful data of sound systems of the first speaking humans exists (by nature, speech leaves no physical trace in its environment), and partly because the questions are simply very difficult. Because the mechanisms involved are bound to be complex and involve the interaction of many environmental, physical, neuro-cognitive and genetic entities, and because data are scarce, computational models have been increasingly used in the past 10 years in the field (Hurford and al., 1998). Indeed, their nature allows on the one hand to test the operational plausibility and feasibility of otherwise highly speculative theories, and on the other hand to gain new insights about how certain aspects can be explained by the intricate dynamics of the complex systems involved.

The research presented here concerns a computational model of the origins of syllable systems, which are thought to be a fundamental unit of the complex sound system of nowadays human languages. It aims at being a plausible implementation, and hence a proof of feasibility, of the theory that claims that sound systems originated and have properties explained by the self-organization of motor, perceptual, cognitive, social and functional constraints that are not linguistically specific, and this in a cultural manner (Steels, 1998). In brief, this theory states that speech is a complex cultural adaptive system. Among the forces at stake are articulatory ease, perceptual distinctiveness, time and memory limitations, lexicon pressure, efficiency of communication, noise in the environment and conformance to the group. The word constraint is used in its most general meaning: it can be obstacle or opportunity as we will see.

A number of computational models concerning the origins of sound systems have already been developed, mainly for phonemes, and more precisely at the vowel level. Two of them are representative: the first one was developed by Lindblom (1992), and consisted in showing that the numerical optimization of a number of motor and perceptual constraints defined analytically allowed to predict the most frequent vowel systems in the human languages (in particular the high occurrence of the 5 vowel system /e i a u o/). Whereas it gave an idea of why vowel systems have the properties observed by phoneticians, it did not give any idea of what process could have achieved this optimization. Indeed, it is not plausible that primitive humans may have willingly computed all possible vowel systems and took an optimal one (and still if this was the case, this does not tell us how they agreed on which of the many solutions that exist). This shortcoming was corrected by the model of de Boer (1999)

, who places itself in a broader class of models consisting in setting up a society of agents endowed with realistic capabilities and physical constraints (whose action on the sound system is not explicit) and have them interact in order to build culturally an efficient communication system (Steels, 1998; Steels and Oudeyer 2000). More specifically, in his model no explicit optimization was performed, but rather near-optimal systems were obtained only as a side effect of adaptation to the task of building a communication system. Coherence did not come from a genetically pre-specified plan, but from the self-organization arising from positive feedback loops.

Much fewer existing models tackle the question of the origins of complex sounds, in particular syllables. Lindblom (1992) and then Redford (1998) have developed models resembling the Lindblom model for vowels: they consist in defining explicitly with analytical formulas a number of constraints, then running an optimization algorithm and showing that near-optimal systems have regularities characteristic of the most common syllable systems in the human languages. An example of regularity is the sonority hierarchy principle, which states that the sonority of syllables tends to increase until their nucleus, and then decrease. The present model aims at applying the multi-agent based modeling paradigm mentioned earlier to the question of the origins and properties of syllable systems: like de Boer's model, it should not only try to explain why syllables tend to be the way they are, but also what actual process built them. An additional requirement needed by this model is the fact that agents should be as realistic as possible, and should operate in the real world. One of the reasons for the need of realism is that previous models have shown that constraints are important to the shape of sound systems: when dealing with too abstract constraints, there is a danger to find wrong explanations. Furthermore, Redford showed that certain phenomena can be understood only by considering the interactions between constraints, so models should try to incorporate most of them.

The next section presents an overview of the model with its different modules. A more comprehensive description can be found in a companion paper (Oudeyer 2001) dedicated to the technical details of the setup. Then we present results about the behavior of the system, and discuss implications for phonetics and phonology, and more generally language epigenesis.

The model

The imitation game

Central to the model is the way agents interact. We use here the concept of game, recently operationalized in a number of computational models of the origins of language (Steels, 1998). A game is a sort of protocol that describes the outline of a conversation, allowing agents to coordinate by knowing who should try to say what kind of things at a particular moment. Here we use the "imitation game" developed by de Boer for his experiments on the emergence of vowel systems.

A round of a game involves two agents, one being called the speaker, and the other the hearer. Here we just note that each possess a repertoire of items/syllables/prototypes, with a score associated to each of them (this is the categorical memory described below). The speaker initiates the conversation by picking up one item in its repertoire and utters it. Then the hearer tries to imitate this sound by producing the item in its repertoire that matches best with what he heard. Then the speaker evaluates whether the imitation was good or not by checking whether the best match to this imitation in its repertoire corresponds to the item uttered initially. Then he gives a feedback signal to the hearer in a non-linguistic manner (see Steels, 1998). Finally, each agent updates its repertoire. If the imitation succeeded, the scores of involved items increase. Otherwise, the score of the item used by the speaker decreases and there are 2 possibilities for the hearer: either the score of the prototype used was below a certain threshold, and this item is modified by the agent who tries to find a better one ; or the score was above this threshold, which means that it may not be a good idea to change this item, and a new item is created, as close to the utterance of the speaker as the agent can do given its constraints and knowledge at this time of its life. Regularly the repertoire is cleant by removing the items that have a score too low. Initially, the repertoires of agents are empty. New items are added either by invention, which takes place regularly in response to the need of growing the repertoire, or by learning from others.

The production module

Vocal tract A physical model of the vocal tract is used, based on an implementation of Cook's model (Cook 1989). It consists in modeling the vocal tract together with the nasal tract as a set of tubes that act as filters, into which are sent acoustic waves produced by a model of the glottis and a noise source. There are 8 control parameters for the shape of the vocal tract, used for the production of syllables. Finally, articulators have a certain stiffness and inertia.

Control system The control system is responsible for driving the vocal tract shape parameters given an articulatory program, which is the articulatory specification of the syllable. Here we consider the syllable from the point of view of the frame-content theory (MacNeilage 1998) which defines it as an oscillation of the jaw (the frame) modulated by intermediary specific articulatory configurations, which represent a segmental content (the content) corresponding to what one may call phonemes. A very important aspect of syllables is that they are not a mere sequencing of segments by juxtaposition: co-articulation takes place, which means that each segment is influenced by its neighbors. This is crucial because it determines which syllables are difficult to pronounce and imitate. We model here co-articulation in a way very similar to what is described in (Massaro 1998), where segments are targets in a number of articulatory dimen-

sions. [1] The constraint of jaw oscillation is modeled by a force pulling in the direction of the position the articulators would have if the syllable was a pure frame, which means an oscillation without intermediary targets. This can be viewed as an elastic whose rest position at each time step is the pure frame configuration at this time step. It is motivated by important neuro-scientific evidence whose synthesis can be found in (MacNeilage, 1998). Finally, and crucially, we introduce a notion of articulatory cost, which consists in measuring on the one hand the effort necessary to achieve an articulatory program and on the other hand the difficulty of this articulatory program (how well targets are reached given all the constraints). This cost is used to model the principle of least effort explained in (Lindblom 1992) : easy articulatory programs/syllables tend to be remembered more easily than others. Agents are initially given a set of pre-defined targets that can be thought to come from an imitation game on simple sounds (which means they do not involve movements of the articulators) as described in (de Boer 2000, Steels and Oudeyer 2000). Although the degrees of freedom that we can control here do not correspond exactly to the degrees that are used to define human phonemes, we chose values (see Oudeyer 2001) that allow them to be good metaphors of vowels (V), liquids (C1) and plosives (C2), which mean respectively sonorant, less sonorant, and even less sonorant phonemes (sonority is directly related to the degree of obstruction of the air flow, which means the more articulators are opened, the more they contribute to a high sonority of the phoneme).

The perception module

The ear of agents consists of a model of the cochlea, and in particular the basilar membrane, as described in (Lyon 1997). It provides the successive excitation of this membrane over time. Each excitation trajectory is discretized both over time and frequency: 20 frequency bins are used and a sample is extracted every 10 ms. Next the trajectory is time normalized so as to be of length 25. As a measure of similarity between two perceptual trajectories, we used a technique well-known in the field of speech recognition, dynamic time warping (Sakoe and Chiba 1980). Agents use this measure to compute which item in their memory is the closest. No segmentation into "phonemes" is done in the recognition process: the recognition is done over the complete unsegmented sound. Agents discover which phonemes compose the syllable only after recognition of the syllable and by looking at the articulatory program associated to the matched perceptual trajectory in the exemplar (see below). In brief, phonemes are not relevant for perception, but only for production. This follows a view defended by a number of researchers (Seguy, Dupoux et Mehler 1995) who showed with psychological experiments that the syllable was the primary unit of recognition, and that phoneme recognition came only after.

[1] The difference is that, as described in the companion paper (Oudeyer 2001), we provide a biologically plausible implementation inspired from a number of neuroscientific findings (Bizzi and Mussa-Ivaldi 1991) and that uses techniques developed in the field of behavior-based robotics (Arkin 1999).

The brain module

The knowledge management module of our agents consists of 2 memories of exemplars and a mechanism to shape and use them. A first memory (the inverse mapping memory) consists of a set, limited in size, of exemplars that serve in the imitation process: they represent the skills of agents for this task. Exemplars are associations between articulatory programs and corresponding perceptual trajectories. The second memory (the categorical memory) is in fact a subset of the inverse-mapping memory, to which a score is added to each exemplar. Categorical memory is used to represent the particular sounds that count as categories in the sound system being collectively built by agents (corresponding exemplars are prototypes for categories). It corresponds to the memory of prototypes classically used in the imitation game (de Boer 1999).

Initially, the inverse mapping memory is built through babbling. Agents generate random articulatory programs, execute them with the control module and perceive the produced sound. They store each trial with a probability inverse to the articulatory cost involved (prob=1-cost). The number of exemplars that can be stored in this memory is typically quite limited (in the experiments presented below, there are 100 exemplars whereas the total number of possible syllables is slightly above 12000). So initially the inverse mapping memory is composed of exemplars which tend to be more numerous in zones where the cost is low than in zones where the cost is higher. As far as the categorical memory is concerned, it is initially empty, and will grow through learning and invention.

When an agent hears a sound and wants to imitate it, he first looks up in its categorical memory (if it is not empty) and find the item whose perceptual trajectory is most similar to the one he just heard. Then he executes the associated articulatory program (noise is always added to target values). Now, when the interaction is finished, in any case (either it succeeded or failed), it will try to improve its imitation. To do that, it finds in its inverse mapping memory the item (it) whose perceptual trajectory matches best (it may not be the same as the categorical item). Then it tries through babbling a small number of articulatory variations of this item that do not belong to the memory: each articulatory trial item is a mutated version of it, i.e. one target has been changed or added or deleted. This can be thought of the agent hearing at a point "ble", and having in its memory the closest item being "fle". Then it may try "vle", "fli", or even "ble" if the chance decides so (indeed, not all possible mutations are tried, which models time constraints: here they typically try 10 mutations). The important point is that these mutation trials are not forgotten for the future (some of them may be useless now, but very useful in the future): each of them is remembered with a probabil-

ity inverse to its articulatory cost. Of course, as we have memory limitation, when new items are added to the inverse mapping memory, some others have to be pruned. The strategy chosen here is the least biased: for each new item, a randomly chosen item is also deleted (only the items that belong to categorical memory can not be deleted).

The evolution of inverse mapping memory implied by this mechanism is as follows. Whereas at the beginning items are spread uniformly across "iso-cost" regions, which means skills are both general and imprecise (they have some capacity of imitation of many kinds of sounds, but not very precise), at the end examplars are clustered in certain zones corresponding to the particular sound system of the society of agents, which means skills are both specialized and precise. This is due to the fact that exemplars closest to sounds produced by other agents are differentiated and lead to an increase of exemplars in their local region at the cost of a decrease elsewhere.

Behavior of the model

Efficiency

The first thing one wants to know is simply whether populations of agents manage to develop a sound system of reasonable size and that allows them to communicate (imitations are successful). Figure 1 and 2 show an example of experiment involving 15 agents, with a memory limit on inverse-mapping memory of 100 exemplars, with vocalizations comprising between 2 and 4 targets included among 10 possible ones (which means that at a given moment, one agent never knows more than about 0.8 percent of the syllable space). In figure 1, each point represents the average success in the last 100 games, and on figure 2, each point represents the average size of categorical memory in the population (i.e. the mean number of syllables in agents' repertoires). We see that of course the success is very high right from the start: this is normal since at the beginning agents have basically one or two syllables in their repertoire, which implies that even if an imitation is quite bad in the absolute, it will still get well matched. The challenge is actually to remain at a high success rate while increasing the size of the repertoires. The 2 graphs shows that it is the case. To make these results convincing, the experiments was repeated 20 times (doing it more is rather infeasible since each experiment basically lasts about 2 days), and the average number of syllables and success was measured in the last 1000 games (over a total of 20000 games): 96.9 percent is the mean success and 79.1 is the mean number of categories/syllables.

The fact that the success remains high as the size of repertoire increases can be explained. At the beginning, agents have very few items in their repertoires, so even if their imitations are bad in the absolute, they will be successfully recognized since recognition is done by nearest-neighbours (for example, when 2 agents have only 1 item, no confusion is possible since their is only

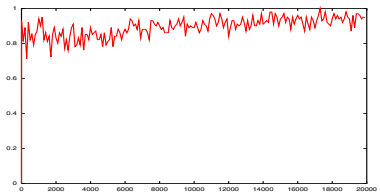


Figure 1: Example of the evolution of success in interactions for a society of agents who build a sound system from scratch

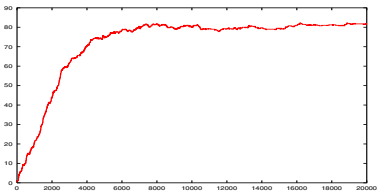


Figure 2: Corresponding evolution of mean number of items/categories in repertoires of agents along with time

1 category !). As time goes on, while their repertoires become larger, their imitation skills are also increasing : indeed, agents explore the articulatory/acoustic mapping locally in areas where they hear other utter sounds, and the new sounds they create are hence also in these areas. The consequence is a positive feed-back loop which makes that agents who knew very different parts of the mapping initially tend to synchronize their knowledge and become expert in the same (small) area (whereas at the beginning they have skills to imitate very different kinds of sounds, but are poor when it becomes to make subtle distinctions in small areas).

This result is relevant to all theories of speech (and more generally, theories of language), innatist or not. Indeed, whereas the literature is rich of reasons explaining why having complex sound systems was an advantage for the first speaking humans is, no precise account of how it could have been built was described. For instance, even Pinker and Bloom (1990), who defend the idea that nowadays humans have a lot of linguistic knowledge already encoded in the genes, acknowledge that it certainly got there through the Baldwin effect and so was initially certainly the result of a cultural process. They give cues about how acquired skills and sound systems could have been transferred into the genes, but not how they got to be acquired from a situation where there was nothing !

Structural properties

Now that we have seen that a communication system was effectively built, one has to look whether the structural properties of the produced repertoires of syllables resemble human syllable systems. Indeed, human syllable systems are far from random: only very few combinations of types of phonemes occur in human languages compared to the high number of mathematically possible ones, and some occur significantly more often than others (Vennemann 1988) . For instance, all languages have CV sylla-

Table 1: % of syllable types in produced and random systems

CV	CVC	CC
25.3/0.2	20.1/1.3	16.1/0.5
CCV	CVVC/CCVC/CVCC	other
14.4/1.3	14.1/22.5	10/74.2

Table 2: % of syllables respecting the sonority hierarchy

full model	jaw constraint removed	chance
70.9 per	21.5 per	5.3 per

bles, but CVCC is rare. The difference in frequencies exist both across and within languages. A first study about the syllable types of the produced systems was achieved. Statistics about the set of all the syllables produced by 20 runs were computed (for each run, measures were done after 20000 games). Table 1 sums up the result by giving the relative frequency in use of a number of syllable types (C means "C1 or C2"). "Relative frequency in use" means that each syllable counts as the number of times it has been used by the agents in the games it played in its life. This is a better measure than simply counting the frequency of occurrence in a syllable system, because it takes into account the fact that certain syllables tend to be adopted earlier than others, which implies that they are used more times than others, and models the relative frequency effects observed within languages. The second percentage measures the proportion of the particular type of syllable in the space of all combinatorially possible ones in the experiment. This can be viewed as a measure of syllable frequencies for randomly generated repertoires.

The first observation we can do is that there is a strong difference between the relative frequencies of syllables in actual systems and in randomly generated systems. Moreover, we find that the ordering between syllables types along their frequency is very similar to the one observed in human languages (Venemann 1988), except for the presence of CC in third position (which we think is due to too low acoustic noise, unlike in the real world). These results are rather consistent with those found by Redford, and conform to what she calls the "iterative principle of syllable structure": "simpler syllables types are expected to occur more frequently than complex ones in a systematic fashion", where the notion of "simplicity" is constructed over the most simple syllable CV: increase in complexity comes by adding C or V iteratively at the end or beginning or CV, and then after by replacing some C by V or the contrary.

A second important tendency of human languages is the "sonority hierarchy principle" [2] Whereas the measures in table 1 indicate that this seems to be the case in our experiment, they are too loose to conclude, especially because they blended C1 and C2. So we made measures over 20 runs about which proportion of syllables belonging to the repertoire of agents did obey the sonority hierarchy principle, using the fact that sonority of V is higher than sonority of C1, which is higher than sonority of C2 (due to the way they obstrue the air flow). Additionally, we made an experiment in which the oscillation of the jaw constraint was removed, in order to evaluate the hypothesis of Peter McNeilage that says that it is the main explanation for the sonority hierarchy principle. Table 1 sums up the results, with a column showing what is the proportion of syllable in the set of combinatorially possible syllables that respect this hierarchy.

We see that the sonority hierarchy is respected by most of the syllables of the emergent repertoires in the standard model. Yet, not all of them respect it, which is not that surprising since syllables like C1C1V do not imply an important deformation of the pure frame and so have a low cost, and do not respect the principle (there are 2 adjacent segments with the same sonority). Anyway, the actual percentage as compared to chance is much higher. When we remove the jaw constraint, we observe that the percentage of syllables respecting this hierarchy drops to around 20 percent, but is still substantially above chance. It indicates that the jaw constraint is crucial, but not the only responsible. In fact, when we remove the jaw constraint, we still start every syllable with the rest position corresponding to the closed jaw. So for instance syllables beginning with a vowel will still have a high articulatory cost. Of course for example C2C2 syllables will have a much lower cost in this case that in the case with jaw oscillation, but these syllables are very sensible to noise and do not have a high perceptual discriminability, which makes agents prune them quite often. As a result, a reasonable proportion of syllables that respect the hierarchy remain.

Until now, we have only looked at how the model produced syllable systems that reflect universal tendencies of human languages. We also have to look how well it matches with the diversity that exists across languages (Vennemann, 1988). Indeed, tendencies are just tendencies and there are cases of languages whose syllable systems properties significantly differ from the mean (for example, in Berber, there are many syllables with long consonant sequences, and more strikingly, there are syllables that do not contain any vowel). Additionally, two languages that have for instance the same relative ratios of syllable types may implement these in very different manners. The first kind of diversity was difficult to observe in a statistically significant manner, since the relative frequencies of syllable types most often are very close to the mean above mentioned, and since not enough experiments were conducted to study rare outliers. Nevertheless, they were observed in a number of particular cases: for example, one of the obtained population

[2] in a syllable, the sonority or loudness first increases to a peak and then may decrease again. It is very rare that for instance it first decreases and then increases or that more than 1 change in sonority direction occurs in one syllable. For instance, "ble" is preferred to "lbe". Sonority/loudness is directly linked to the degree of obstruction of the air, and in particular to the degree of opening of the jaw.

had 55 percent of CVC/CCV syllables against only 20 percent of CV syllables. Some categorical differences were also observed: several populations did not have any CVVC or CVCC syllables for instance. The second kind of diversity was easier to observe in the system: you never get the same repertoires in 2 different runs of the experiment. In the 20 runs used for the experiments above, the mean number of common syllables was 20.2 (repertoires had sizes varying between 70 and 88), among which mainly 2-phonemes syllables due to their small number. Of course this result is not directly transposable to real languages since we always gave here the same set of phonemes in the beginning, whereas in reality these phonemes are not pre-given but should co-evolve with syllables, and so may lead to repertoire of syllables composed of very different phonemes [3]. Nonetheless, we get a good idea of how universal tendencies come from the fact that there are non-linguistically specific constraints/biases in the problem that agents are solving, whereas diversity comes from both the fact that these constraints are soft and that there exist many satisfying solutions to the problem. Operationally speaking, variety emerges because there is stochasticity locally in time and space, which makes that different societies may engage different pathways due to historical events: indeed, historicity is fundamental to the explanation of diversity. This view contrasts in different aspects with a number of innatist theories, especially optimality theory (Archangeli and Lagendoen 1997). Of course, there is a common point with optimality theory at a very general level: constraints are crucial to the explanation of language universals and diversity. Yet, a fundamental difference is the nature of constraints: in the case of optimality theory, they are linguistically specific, whereas here they are generic constraints of the motor, perceptual and cognitive apparati (we also have social constraints that are far from any concept in OT) [4]. Now, the second important difference is the way these constraints are used to explain diversity: in OT, a particular syllable system corresponds to a particular ordering of constraints (some are stronger than others, which means that a low ranked constraint may be over-ridden if one has to satisfy a higher ranked constraint), which means a different constraint satisfaction problem. Conversely, in OT, one ordering of constraint implies a fixed syllable system (in terms of syllables types). On the contrary, here we do not require a different set of constraints to obtain different kinds of systems, because there are many syllables systems that can be developed and allow efficient communication given only one set of constraints. Our model thus avoids a number of theoretical problems that OT is faced with: Where do the linguistic constraints come from ? If they are in the genes, how did they get there ? Why are there different orderings of constraints ? How one can pass from a set of constraints to another (which must happen since language evolves and syllable systems change) ?

[3] This is a limit of the model (that the model of Redford has also), but we think this limitation was necessary as a first step so that the resulting dynamics would not get too complicated to analyse.

[4] An example of constraint in OT is the *COMPLEX constraint which states that syllables can have at most one consonant at an edge or the NOCODA constraint which says that syllables must end with vowels.

Conclusion

We have presented an operational model of the origins of syllable systems whose particularity is the stress on embodiment and situatedness constraints/opportunities, which implies the avoidance of many shortcuts usually taken in the literature. It illustrates in details (and brings more plausibility) the theory which states that speech originated in a cultural self-organized manner, taking as a starting point a set of generic non-linguistically specific learning, motor and perceptual capabilities. In addition to the demonstration of how an efficient communication system could be built with this parsimonious starting point, some specific properties that are known about human sound systems can be explained by our model : on the one hand, universal tendencies like the preference for CV and CVC syllable types and the sonority hierarchy principle ; on the other hand, diversity. A forthcoming paper will present other properties of human sound systems predicted by this model, among which the critical period phenomenon, the difficulty to learn a second language and the difficulty to learn artificial random sound systems as compared to “natural ones”.

References

Altman, (1995), Cognitive Models of Speech Processing, Psycholinguistics and Computational Perspectives, MIT Press.

Archangeli D., Langendoen T. (1997) Optimality theory, an overview, Blackwell Pulishers.

Arkin, R. (1999) Behavior-based Robotics, MIT Press.

Bizzi E., Mussa-Ivaldi F., Giszter S. (1991) Computations underlying the execution of movement, Science, vol. 253, pp. 287-291.

de Boer, B. (1999) Investigating the Emergence of Speech Sounds. In: Dean, T. (ed.) Proceedings of IJCAI 99. Morgan Kauffman, San Francisco. pp. 364-369.

Chomsky, N. and M. Halle (1968) The Sound Pattern of English. Harper Row, New york.

P. R. Cook, ”Synthesis of the Singing Voice Using a Physically Parameterized Model of the Human Vocal Tract,” Proc. of the International Computer Music Conference, pp. 69-72, Columbus, OH, 1989.

Hurford, J., Studdert-Kennedy M., Knight C. (1998), Approaches to the evolution of language, Cambridge, Cambridge University Press.

Lindblom, B. (1992) Phonological Units as Adaptive Emergents of Lexical Development, in Ferguson, Menn, Stoel-Gammon (eds.) Phonological Development: Models, Research, Implications, York Press, Timonnium, MD, pp. 565-604.

Lyon, R. (1997), All pole models of auditory filtering, in Lewis et al. (eds.) Diversity in auditory mechanics, World Scientific Publishing.

Massaro, D. (1998) Perceiving talking faces, MIT Press.

MacNeilage, P.F. (1998) The Frame/Content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21, 499-548.

Oudeyer P-y. (2001) The origins of syllable systems in a society of truly autonomous robots, submitted to Artificial Intelligence Journal.

Oudeyer P-y. (2001) Coupled Neural Maps for the Origins of Vowel Systems, to appear in the proceedings of ICANN’01, International Conference on Artificial Neural Networks, Springer Verlag.

Pinker, S., Bloom P., (1990), Natural Language and Natural Selection, The Brain and Behavioral Sciences, 13, pp. 707-784.

Redford, M.A., C. Chen, and R. Miikkulainen (1998) Modeling the Emergence of Syllable Systems. In: Proceedings of the Twentieth Annual Conference of the Cognitive Science Society. Erlabum Ass. Hillsdale.

Sakoe H., Dynamic programming optimization for spoken word recognition, IEEE Transaction Acoustic., Speech, Signal Processing, vol. 26, pp. 263-266.

Segui, J., Dupoux E., Mehler J. (1995) The role of the syllable in speech segmentation, phoneme identification, and lexical access, in Altman, (ed.), Cognitive Models of Speech Processing, Psycholinguistics and Computational Perspectives, MIT Press.

Steels, (1998), Synthesizing the origins of language and meaning using co-evolution, self-organization and level formation, in Hurford, Studdert-Kennedy, Knight (eds.), Cambridge University Press, pp. 384-404.

Steels L., Oudeyer P-y. (2000) The cultural evolution of syntactic constraints in phonology, in Bedau, McCaskill, Packard and Rasmussen (eds.), Proceedings of the 7th International Conference on Artificial Life, pp. 382-391, MIT Press.

Vennemann, T. (1988), Preference Laws for Syllable Structure, Berlin: Mouton de Gruyter.

Prototype Abstraction in Category Learning?

Thomas J. Palmeri (thomas.j.palmeri@vanderbilt.edu)
Marci A. Flanery (marci.flanery@vanderbilt.edu)
Department of Psychology; Vanderbilt University
Nashville, TN 37240 USA

Abstract

Do people learn categories by abstracting prototypes, forming simple rules, remembering specific exemplars, or by some combination of these? Although some consensus seems to be emerging for a combination of rule formation and exemplar memorization, recent work has revived interest in prototype abstraction (e.g., Smith et al., 1997; Smith & Minda, 1998). We reexamined this recent evidence with an eye toward an alternative simple strategy subjects could use within those particular studies. A very simple strategy, available in some categorization tasks in which corrective feedback is supplied, is to classify the current stimulus in the same category as the previous stimulus if the two are sufficiently similar to one another. This simple strategy makes no recourse to stored category representations of any kind. And this strategy will be useful only under certain circumstances. Reexamining the work by Smith and colleagues, we found that those category structures that produced evidence for prototype abstraction could be "learned," at least to some degree, using this simple strategy. Moreover, simulated data sets created using this simple strategy were better fitted by a prototype model than an exemplar model. We argue that evidence for prototype abstraction from the studies by Smith and colleagues may be weaker than they originally claimed.

Introduction

Do people learn categories by abstracting prototypes, forming rules, remembering exemplars, or some combination of these? In the domain of learning novel perceptual categories, we can trace an evolving dominance of various theoretical accounts from rule formation via hypothesis testing in the early years of categorization research (e.g., Bruner et al., 1956; Trabasso & Bower, 1968), to prototype abstraction (e.g., Homa, 1984; Posner & Keele, 1968; Reed, 1972), to exemplar storage and retrieval (e.g., Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1986). More recently, there has been a reemerging interest in the importance of rule learning in categorization (e.g., Allen & Brooks, 1991; Nosofsky et al., 1994; Nosofsky & Palmeri, 1998). This has led to a variety of hybrid accounts proposing a combination of rule learning and exemplar memorization (e.g., Erickson & Kruschke, 1998; Johansen & Palmeri, 2001; Palmeri, 1997; Palmeri & Johansen, 1999; see also Ashby et al., 1998). Arguably, there seems to be an emerging consensus for some kind of combination of rule formation and exemplar memorization in category learning. However, there has also been reemerging interest in the potential role of prototype abstraction in category learning as well, at least under certain conditions (e.g., Smith et al., 1997; Smith & Minda, 1998). The goal of the present article was to critically reexamine this evidence for prototype abstraction that has been provided by Smith and colleagues.

Evidence for Prototype Abstraction

According to prototype models, people learn categories by averaging their experiences with specific exemplars to derive an abstract prototype and classify new objects according to similarity to stored prototypes. By contrast, according to exemplar models, people remember information about specific exemplars, with no abstraction of prototypes or other summary representations, and classify new objects according to their similarity to the stored category exemplars. Numerous studies have compared and contrasted the ability of exemplar models and prototype models to account for observed categorization data – the majority of studies found that exemplar models provided a far superior account, both qualitatively and quantitatively (e.g., Buse-

Table 1: An example category structure from Smith and Minda (1998), Experiments 1 and 2.

Category A		Category B	
Structure	Stimuli	Structure	Stimuli
Categories Linearly Separable			
000000	banuly	111111	kepiro
010000	benuly	111101	kepib
100000	kanuly	110111	keniro
000101	banib	101110	kapiry
100001	kanub	011110	bepiry
001010	bapury	101011	kapuro
011000	bepuly	010111	beniro
Categories Not Linearly Separable			
000000	gafuzi	111111	wysero
100000	wafuzi	011111	gysero
010000	gyfuzi	101111	wasero
001000	gasuzi	110111	wyfero
000010	gafuri	111011	wysuro
000001	gafuzo	111110	wyseri
111101	wysezo	000100	gafezi

meyer et al., 1984; Medin & Schaffer, 1978; Palmeri & Nosofsky, 2001; Shin & Nosofsky, 1992).

Challenging this previous work, there have been some recent studies that have reexamined the potential explanatory power of prototype models compared to exemplar models under a variety of conditions. In this work, prototype models have been found to provide superior fits to some data (Smith et al, 1997), especially early in category learning (Smith & Minda, 1998).

Smith et al. (1997) conducted a series of category learning experiments using structures and stimuli quite similar to those shown in Table 1. Stimuli were six-letter (6D) pronounceable nonsense words composed of alternating consonants and vowels. At each position in the word, one of two possible letters could appear (e.g., the first letter could be either g or w). The two categories were generally formed around a prototype (e.g., *gafuzi* versus *wysero*) with most category members differing from their category prototype along one or two dimensions (e.g., *gasuzi*). Category structures were either linearly separable (LS) or nonlinearly separable (NLS). Linearly separable categories are those that can be partitioned on the basis of some weighted additive combination of information along the individual dimensions (Medin & Schwanenflugel, 1981). As shown in Table 1, the nonlinearly separable categories have an exception item that is similar to the prototype of the contrasting category (e.g., *wysezo* in the *gafuzi* category). In the first experiment of Smith et al. (1997), one set of category structures was relatively easy (ELS and ENLS) and another set of category structures was relatively difficult (DLS and DNLS).

Each subject learned one of the four possible category structures. On each trial of the experiment, the subject was presented with one of the items randomly selected from one of the two categories to be learned. The subject classified the item as a member of category A or category B and then received corrective feedback.

Smith et al. (1997) tested the ability of prototype and exemplar models to account for the probabilities of classifying each stimulus as a member of category A or B on an individual subject basis. Across all four conditions of Experiment 1, they found that the prototype model provided a better account of the observed data than the exemplar model for half of their subjects. Their evidence for a prototype model advantage is summarized in Table 2. A YES in the second column signifies that at least some proportion of the individual subjects were displaying categorization behavior that a prototype model was better able to account for.

To illustrate, focusing on the ENLS category, the subgroup of subjects that a prototype model better accounted for averaged 92%, 78%, and 23% correct on the prototypes, normal items, and exception items, respectively. Consistent with the predictions of the prototype model, these subjects erroneously classified the exception items as being members of the category of the prototype they were most similar to. By contrast, the subgroup of subjects that an exemplar model better accounted for averaged 81%, 79%, and 51% correct on the prototypes, normal items, and exceptions.

Smith et al. (1997) and Smith and Minda (1998) provided further evidence for a prototype model advantage across a series of experiments. One manipulation they performed varied the number of dimensions present in the stimuli. With six dimensions (6D) it is possible to create well-differentiated categories (those with high within-category similarity and low between-category similarity) with many members. However, with only four stimulus dimensions (4D), categories tend to be much less differentiated and tend to be much smaller. Smith and colleagues have argued that prototype models show their advantage where prototype abstraction is most likely to succeed, under those conditions where categories are composed of stimuli with many dimensions, where categories are large in size, and where categories are well differentiated. (As a further manipulation, in some experiments nonsense words were used, while in other experiments cartoon animals were used.) As summarized in Table 2, Smith et al. (1997) documented a series of conditions under which some proportion of subjects used prototype abstraction. In Smith and Minda (1998), evidence for prototype abstraction, if present, was observed in the early stages of category learning; exemplar models generally fared better than prototype models in later stages of learning. For experiments from that article, a YES in the second column of Table 2 signifies that a prototype model provided a superior account of early stages of category learning. As shown in Table 2, for some of the category structures Smith and colleagues tested, no evidence for prototype abstraction was observed.

A Simple Categorization Strategy

In all of the experiments cited in Table 2, prototype and exemplar models were tested on their ability to account for category learning data. This data was obtained from trials in which subjects were presented a stimulus, made a response, received feedback, were presented the next stimulus, made a response, received feedback, and so on. Our goal was to investigate whether some subjects could be using some form of the following very simple strategy to provide the correct answer without relying on abstracted prototypes or learned exemplars.

We will use the category structure shown in Table 1 as an example. Suppose on some trial, a subject is shown the following stimulus

gafuzi

and is then asked to classify it as a member of category A or category B. The subject responds *category A* and the computer provides the following feedback

CORRECT, *gafuzi* is a member of *category A*

Suppose the subject is next presented this stimulus

gafuri

and is asked to classify it. The subject could rely on abstracted prototypes, or remembered exemplars, or

formed rules. But perhaps a far simpler strategy is to classify it in the same category as gafuzi since they are so similar to one another. The computer just verified that gafuzi is a member of *category A* so it is reasonable to guess that gafuri might also be a member of *category A* as well. The subject responds *category A* and the computer provides the following feedback

CORRECT, gafuri is a member of *category A*

Suppose the subject is next presented this stimulus

wasero

and is asked to classify it. Well, this stimulus looks quite different from the previous stimulus, gafuri, so it might make sense to classify it in the opposite category as gafuri. The subjects responds *category B* and the computer provides the following feedback

CORRECT, wasero is a member of *category B*

Finally, suppose the subject is presented this stimulus

gafezi

Well, this stimulus looks very different from the previous one, wasero, so it might make sense to classify it in the opposite category. The subject responds *category A* and the computer provides the following feedback

WRONG, gafezi is a member of *category B*

Examining Table 1, we see that this stimulus is the exception to category B. Using this very simple strategy, our subject would seem to perform quite well at classifying everything but the exceptions. Recall that Smith et al. (1997) reported that their subjects whose data was best fit by a prototype model consistently classified the exceptions in the wrong category as well.

Table 2: Evidence for Prototype Abstraction from Smith et al. (1997) and Smith & Minda (1998). See text for a key to the experiment notation.

Experiment	Prototypes?	Strategy?
Smith et al. (1997)		
Experiment 1 6D ELS	YES	YES
Experiment 1 6D DLS	YES	YES
Experiment 1 6D ENLS	YES	YES
Experiment 1 6D DNLS	YES	YES
Experiment 2 4D NLS	NO	NO
Experiment 2 6D NLS	YES	YES
Smith & Minda (1998)		
Experiment 1 6D LS	YES	YES
Experiment 1 6D NLS	YES	YES
Experiment 2 6D LS	YES	YES
Experiment 2 6D NLS	YES	YES
Experiment 3 4D LS	NO	NO
Experiment 3 4D NLS	NO	NO
Experiment 4 4D NLS	NO	NO
Experiment 4 6D NLS	YES	YES

Note. The second column (Prototypes) indicates whether evidence for prototype abstraction was observed. The third column (Strategy) indicates whether the simple strategy yields above chance performance.

Does the Simple Strategy Work?

This is certainly a strategy that subjects could use to classify stimuli in an experiment. But, does it really work? In many cases, no. For example, let us consider the experiments reported by Medin and Schwanenflugel (1981). Different groups of subjects learned linearly separable and nonlinearly separable categories. Their results were important because they found that NLS categories could be easier to learn than LS categories, a result inconsistent with additive prototype models. By contrast, this result was an a priori prediction of multiplicative exemplar models. Let us first examine the category structure from the third experiment from Medin and Schwanenflugel (1981) in some detail. Using an abstract notation, for the LS structure, stimuli in category A were 0111, 1110, and 1001 and stimuli in category B were 1000, 0001, and 0110. For the NLS structure, stimuli in category A were 1100, 0011, and 1111, and stimuli in category B were 0000, 0101, 1010. We performed a Monte Carlo simulation of the simple strategy using these two category structures. For each of 1000 simulated subjects for each structure, we generated a random sequence of stimulus trials. On each simulated trial, if the current stimulus matched the previous one on more than two dimensions, then the same category response as the previous stimulus was used. If the current stimulus matched the previous one on fewer than two dimensions, then the other category response was used. If the current stimulus matched the previous one on exactly two dimensions, then a random response

Table 3: Best fitting model (exemplar or prototype) to data simulated using the simple strategy. See text for a key to the experiment notation.

Experiment	Prototypes?	Model
Smith et al. (1997)		
Experiment 1 6D ELS	YES	Prototype
Experiment 1 6D DLS	YES	Prototype
Experiment 1 6D ENLS	YES	Prototype
Experiment 1 6D DNLS	YES	Prototype
Experiment 2 4D NLS	NO	–
Experiment 2 6D NLS	YES	Prototype
Smith & Minda (1998)		
Experiment 1 6D LS	YES	Prototype
Experiment 1 6D NLS	YES	Prototype
Experiment 2 6D LS	YES	Prototype
Experiment 2 6D NLS	YES	Prototype
Experiment 3 4D LS	NO	–
Experiment 3 4D NLS	NO	–
Experiment 4 4D NLS	NO	–
Experiment 4 6D NLS	YES	Prototype

Note. The second column (Prototypes) indicates whether evidence for prototype abstraction was observed. The third column (Model) indicates whether the Prototype or Exemplar model provided a better fit to the simulated data.

was generated. Averaging across 1000 simulated subjects, this strategy produced just 34.1% accuracy on the LS structure and 33.7% accuracy on the NLS structure. To see why this simple strategy produced accuracy far worse than just guessing, let us examine the NLS structure. Both NLS categories contain stimuli that mismatch each other on every dimension (1100 and 0011 in category A, 0101 and 1010 in category B). When these mismatching stimuli follow one another, they always produce the wrong response (e.g., erroneously categorizing 0011 as a member of category B because it is preceded by 1100 which was labeled a member of category A). Moreover, on other pairs of sequential trials, stimuli that follow one another match on exactly half the dimensions, producing a random response.

This simple strategy fails at other category structures as well. For the second experiment of Medin and Schwanenflugel (1981), the simple strategy produced 46.6% accuracy for their LS structure. For the category structure from Experiment 4 of Medin and Schaffer (1978), the simple strategy produced 44.6% accuracy. Applying the simple strategy to the classic category structures from Shepard, Hovland, and Jenkins (1961), we obtain predicted accuracies for their problem Types I-IV as following: Type I: 70.8%, Type II: 42.7%, Type III: 57.3%, Type IV: 57.0%, Type V: 43.1%, Type VI: 15.2%. In addition to underpredicting the overall level of accuracy observed when subjects learn these various categories, this simple strategy mispredicts the order of difficulty of the various problem types. For separable-dimension stimuli, the difficulty of the problems is ordered I < II < III,IV,V < VI (Nosofsky et al., 1994; see also Nosofsky & Palmeri, 1996). Clearly, this simple strategy is not what subjects can use in many categorization experiments.

But, the simple strategy does work well when "learning" other category structures. Let us turn now to the category structures shown in Table 1, which were used in Experiments 1 and 2 of Smith and Minda (1998). Following the procedure described above, we used a Monte Carlo simulation procedure to generate categorization responses for the LS and NLS categories using the simple strategy. For each of 1000 simulated subjects, we generated a random sequence of stimuli, with each stimulus presented once per block. On each trial, if the current stimulus matched the previous one on more than three dimensions, then the category of the previous stimulus was used. If the current stimulus matched the previous one on fewer than three dimensions, then the other category was used. If the current stimulus matched the previous one on exactly three dimensions, then a random response was generated.

Using this simple strategy, accuracy of approximately 73% correct was possible for both the LS and NLS category structures (excluding the two exceptions in the NLS structure, which the strategy erroneously classified, the overall accuracy for the remaining items was over 84%). The overall performance is less than the accuracies achieved by subjects in Smith and Minda (1998) by the end of learning, which was slightly over 80% correct for both structures. However, in their experiment, evidence for the use of prototypes was only observed during the early blocks of learning. Smith and Minda fitted the prototype and exemplar models to blocks of 56 trials and found that the prototype model fitted better than the exemplar model during the early blocks of learning. It seems quite possible that subjects might start out using the simple strategy during the early blocks of learning, especially since the strategy correctly classifies nearly three out of four items. As subjects acquire more experience with the categories, they may begin to shift to using stored exemplar information to improve performance.

We next tested the ability of this simple strategy to correctly categorize stimuli from the other category structures used by Smith and colleagues. As summarized in Table 2, the simple strategy produced above chance categorization in just those category structures that Smith and colleagues found evidence for prototype abstraction. For notation, those category structures for which the simple strategy works are indicated by a YES in the third column of Table 2. Could the apparent use of prototypes actually be a signature for the use of this very simple strategy instead?

Which Model Fits Better?

Suppose that subjects are engaging in this simple strategy of comparing the current stimulus with the previous one and selecting the category label accordingly. Let us further assume that they are not abstracting prototypes, are not learning rules, and are not remembering exemplars. Smith and colleagues obtained data from their subjects and compared how well a prototype model and an exemplar model accounted for categorization judgments. If subjects are using the simple strategy, would a prototype model or an exemplar model provide a better account of the categorization judgments produced using this simple strategy?

We will focus on just those structures for which the simple strategy actually yields above chance categorization, as indicated by a YES in the third column of Table 2. Using the simple strategy, we employed the Monte Carlo simulation techniques discussed earlier to generate data from a large number of simulated subjects for each of the indicated category structures. We then examined how well a prototype model or an exemplar model accounted for this simulated data. Clearly one possibility is that the prototype model accounted for some of the simulated data and an exemplar model accounted for the rest of the simulated data. A far more interesting possibility is that either an exemplar model or a prototype model provided a better account for the entire set of simulated data. This would pose an interesting problem of identifiability. The data were generated using a simple strategy of local stimulus comparisons without storing long-term category representations of

any kind. Yet by comparing just a prototype model and an exemplar model, we may erroneously conclude on the basis of model fits that subjects were actually abstracting prototypes or remembering exemplars.

To be specific, we fitted a prototype model and an exemplar model to the simulated data generated using the simple strategy. For the exemplar model, an item to be classified is compared with the stored exemplars of category A and category B. The probability of classifying the item into one of those categories is given by the relative summed similarity of the item to the stored exemplars of the two categories. For the prototype model, the probability of classifying the item is given by the relative similarity to the prototypes of the two categories. Similarity between an item i and a stored exemplar j (or an abstracted prototype j) is given by

$$S_{ij} = \prod_m s_m^{\delta(i,j)}$$

where the $0<s_m<1$ are free parameters along all m dimensions. The $\delta(i,j)$ is a function that returns a 0 if i and j match along dimension m and returns a 1 if i and j mismatch. A small value of s_m indicates that dimension m is particularly diagnostic. Because of the multiplicative similarity rule, mismatches along that dimension will have a large effect on decreasing similarity.

For the exemplar model, evidence for category A, E_A, is found by summing the similarities of an item to all exemplars in category A, and evidence for category B, E_B, is found by summing the similarities to all exemplars in category B. For the prototype model, evidence for category A, E_A, is the similarity to the category A prototype, and the evidence for category B, E_B, is the similarity to the category B prototype. The probability of classifying i into category A is then given by

$$P(A \mid i) = \frac{b_A E_A}{b_A E_A + b_B E_B}$$

where b_A is the category A bias (as might be expected, the bias terms did not contribute to any significantly improved fits to the simulated data). The prototype model and the exemplar model were fitted to the simulated data using a hill-climbing procedure that located parameters that minimized the sum of squared error between simulated observations and model predictions.

The summary of this modeling was straightforward. The prototype model provided a better account of the data that was simulated using the simple strategy than did the exemplar model. This finding is summarized in Table 3. As shown in the third column, for every structure which could be "learned" using the simple strategy, and for which Smith and colleagues found evidence for prototype abstraction, the prototype model provided a superior account of the simulated data. Although the data was generated using a strategy that made no recourse to abstract prototypes, the prototype model fitted that simulated data better than the exemplar model. If subjects were indeed using this simple strategy, one might erroneously conclude that they were abstracting prototypes when in fact they were relying on local stimulus information to make a categorization decision.

Summary and Conclusions

Smith and colleagues (Smith et al., 1997; Smith & Minda, 1998) have provided evidence for prototype abstraction in category learning. We noted that in their experiments, they examined data from learning trials in which feedback was always provided. We investigated the possibility that subjects could use this corrective feedback to classify the subsequent item in the category learning task without making recourse to long-term category representations of any kind. According to this simple strategy, subjects must only compare the current item with the previous one. The previous item had been labeled explicitly by the experimenter in the form of corrective category feedback. The current item is classified in the same category as the previous one if they are sufficiently similar to one another, otherwise the current item is classified in the other category.

First, we observed that those structures that Smith and colleagues found evidence for prototype abstraction were those category structures which could be "learned" using this simple strategy. In other words, engaging in these local stimulus comparisons could produce categorization accuracy greater than chance. For comparison, we documented a number of other category structures for which this simple strategy would be unsuccessful.

Second, we simulated data using this simple strategy for those structures for which the simple strategy would work. We then fitted a prototype and an exemplar model to this simulated data. In every case, the prototype model fitted the simulated data better. If subjects were to use this simple strategy, without relying on stored category representations, we might erroneously conclude from these model fits that subjects were abstracting prototypes when they were not.

In a recent paper, Stewart, Brown, and Chater (in press) documented evidence for the use of sequential information in categorization similar to what we are proposing. Using what they called a memory and contrast (MAC) strategy, they tested whether subjects would respond with the same category as on the previous trial if there was a small difference between the two stimuli, and a different label if the difference was large, just like the simple strategy we investigated. Not only did they demonstrate that the MAC strategy could achieve well over 80% accuracy on the category structures they tested, but they also reported highly systematic sequence effects in the experiments they reported, which were consistent with the use of a MAC strategy. The sequence effects they observed were inconsistent with exemplar models and other models they investigated.

So, the conclusion of our work along with that of Stewart and colleagues is perhaps best described as a cautionary tale. When we engage subjects in category learning experiments, our goal is typically to understand something about the long-term category representations

that subjects may have formed about those categories. Yet, subjects will use whatever information they have available to them to make a correct response, perhaps without even using any long-term category knowledge. They can clearly use feedback from previous trials to categorize on a current trial, at least under some circumstances. Or they can learn something about categories during testing in ways that may be entirely unanticipated by the investigator.

Acknowledgments

This work was supported by Vanderbilt University Research Council Grants, Grant BCS-9910756 from the National Science Foundation, and Grant 1R01 MH61370 from the National Institute of Mental Health.

References

Allen, S.W., & Brooks, L.R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General, 120*, 3-19.

Ashby, F.G., Alfonso-Reese, L.A., Turken, A.U., & Waldron, E.M. (1998). A formal neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*, 442-481.

Bruner, J.S., Goodnow, J.J., & Austin, G.A. (1956). *A study of thinking*. New York: Wiley.

Busemeyer, J.R., Dewey, G.I., & Medin, D.L. (1984). Evaluation of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 638-648.

Erickson, M.A., & Kruschke, J.K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127*, 107-140.

Hintzman, D.L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*, 411-428.

Homa, D. (1984). On the nature of categories. *Psychology of Learning and Motivation, 18*, 49-94.

Johansen, M.K., & Palmeri, T.J. (2001). Representational shifts in category learning. *Manuscript under review*.

Medin, D.L., & Schaffer, M.M. (1978). A context theory of classification learning. *Psychological Review, 85*, 207-238.

Medin, D.L., & Schwanenflugel, P.J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 75*, 355-368.

Nosofsky, R.M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.

Nosofsky, R.M., Gluck, M., Palmeri, T.J., McKinley, S.C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition, 22*, 352-369.

Nosofsky, R.M., & Palmeri, T.J. (1996). Learning to classify integral-dimension stimuli. *Psychonomic Bulletin & Review, 3*, 222-226.

Nosofsky, R.M., Palmeri, T.J., & McKinley, S.C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101*, 53-79.

Nosofsky, R.M., & Palmeri, T.J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*, 266-300.

Nosofsky, R.M., & Palmeri, T.J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review, 5*, 345-369.

Palmeri, T.J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 324-354.

Palmeri, T.J., & Johansen, M.K. (1999). Prototypes, rules, and instances in category learning. *Abstracts of the Psychonomic Society: 40th Annual Meeting, 4*, 98.

Palmeri, T.J., & Nosofsky, R.M. (2001). Central tendencies, extreme points, and prototype enhancement effects in ill-defined perceptual categorization. *The Quarterly Journal of Experimental Psychology, 54*, 197-235.

Posner, M.I., & Keele, S.W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353-363.

Reed, S.K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*, 382-407.

Shepard, R.N., Hovland, C.L., & Jenkins, H.M. (1961). *Psychological Monographs, 75* (13, Whole No. 517).

Shin, H.J., & Nosofsky, R.M. (1992). Similarity-scaling studies of dot-pattern classification and recognition. *Journal of Experimental Psychology: General, 121*, 278-304.

Smith, J.D., Murray, M.J., & Minda, J.P. (1997). Straight talk about linear separability. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 659-680.

Smith, J.D. & Minda, J.P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1411-1436.

Stewart, N., Brown, G.D.A., & Chater, N. (in press). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Trabasso, T., & Bower, G.H. (1968). *Attention in learning: Theory and research*. New York: Wiley.

The Role of Velocity in Affect Discrimination

Helena M. Paterson (helena@psy.gla.ac.uk)
Glasgow, G12 8QB

Frank E. Pollick (frank@psy.gla.ac.uk)
Department of Psychology, 58 Hillhead Street,
Glasgow, G12 8QB

Anthony J. Sanford (tony@psy.gla.ac.uk)
Department of Psychology, 58 Hillhead Street,
Glasgow, G12 8QB

Abstract

Two experiments are described that examine the role of speed in the categorisation of affective biological motion displays. For the first experiment movements were recorded for 10 affects and the point-light animations of them were shown to participants in a recognition task. The resultant confusion matrices were analysed using the ALSCAL multi-dimensional scaling procedure and produced a 2-dimensional psychological space. The psychological space for discrimination was similar to that from recent models of experienced affect in that the first dimension corresponded to the activation dimension from these models. A strong correlation between the movement speed and the activation dimension confirmed the finding. From these results it would appear that the mapping between stimulus properties and representation of activation in affect is a fairly direct one. For the second experiment more sad, angry and neutral movements were collected. New movements of different duration, but identical spatial displacement were made using an interpolation algorithm. Observers viewed the movements as point light displays and heir task was to rate intensity of affect. Results from this experiment indicate speed plays a major role in modulating the intensity of activation in perceived affect.

Introduction

Humans can easily tell each other apart and interpret subtle differences in behaviour that communicates intentions, identity and emotions easily. Cues from facial features are used for much of this recognition, however, highly impoverished stimuli - such as point-light displays - convey sufficient information for the recognition of such person properties (Barclay, Cutting & Kozlowski, 1978, Cutting & Kozlowski, 1977; Dittrich, Troscianko, Lea & Morgan, 1986; Hill & Pollick, 2000; Kozlowski & Cutting, 1978; Mather & Murdoch, 1994; Runeson & Frykholm, 1981; Runeson & Frykholm, 1983; Walk & Homan, 1984). The cues that convey this information in biological motion are of primary interest to us and using point light displays of human arm movements, we have concentrated on the recognition of emotion from biological motion.

In exploring the way in which humans recognise affect, it is possible not only to look at the accuracy with which an affect is recognised, but also at the structure of the representation of affect. A number of models for the structure of experienced affect have been suggested that resemble each other in a number of factors (Russell, 1980; Watson and Tellegen, 1985; Thayer, 1989, Larsen and Diener; 1992; Feldman, Barrett and Russell, 1999). The similarities between these models are that the structure of affect is a two-dimensional and continuous structure. This structure is referred to as a circumplex model (Feldman, Barrett and Russell, 1999). The two dimensions of the circumplex models are bipolar and independent. One dimension represents valence (for instance hedonic tone, pleasant – unpleasant) and the other, arousal or activation (arousal – sleep/ activated – deactivated). The models are also continuous, with affects falling on a circle, centred on the origin of the psychological space defined by the two dimensions.

Although these models were established as representing one's own experience of affect, there is recent evident to suggest that experience and perception interact when observing another person's actions (Decety and Grezes, 1999; Rizzolatti, Fadiga, Gallese and Foggassi, 1996). Additionally biological motion relies on both specialised bottom-up processes of motion detection (Mather, Radford and West, 1992; Neri, Morrone and Burr, 1998) and interactions between these and top-down processes (Shiffrar and Freyd, 1990, 1993; Thornton, Pinto and Shiffrar, 1998). In the case of affect perception such a top-down process may well originate from the influence of an internal structure of affect.

It seems reasonable, therefore, to further explore the possible relationship between perception and structure of affect. There is also, currently, little research that concentrates specifically at the recognition of emotion

from biological motion. The special case of interpreting stylised dance movements from point-light displays, has received some attention (Dittrich, Troscianko, Lea & Morgan, 1986, Walk & Homan, 1984). Dittrich et al found good evidence that these movements can be expressive, however, they did not address more typical movements.

Two experiments are reported that highlight the relationship between the structure of experienced affect and perception of other's affect from simple arm movements. We also report the importance of the role that the speed of such movements, play in affect discrimination.

Experiments

General Methods

Movement Collection Arm movement data was obtained using a three-dimensional position measurement system (Optotrak, Northern Digital). Actors read the emotional scene setting story and then performed drinking and knocking actions. While they made the movements, the position of their head, right shoulder, elbow, wrist, and the first and fourth metacarpal joints were recorded using infra red emitting diodes.

Each movement record was processed to obtain the start and end of the movement as well as other kinematic properties such as tangential velocity, acceleration and jerk of the wrist. The start of the movement was defined as the point 116 msec before the tangential velocity of the wrist rose above a criterion value, and the end by the point 116 msec after the velocity passed below the criterion. This start/end velocity criterion was defined as 5% of the peak tangential velocity of the wrist. To measure kinematics, instantaneous measures of the wrist kinematics (velocity, acceleration and jerk) were taken and kinematic markers of duration, average velocity, peak velocity, peak acceleration, peak deceleration and jerk index were identified. Jerk index was defined as the magnitude of the jerk averaged over the entire movement and relates to the smoothness of a movement (Flash & Hogan, 1985).

Stimuli In all experiments each recorded point of the arm movement was presented as a point light on a graphics computer (Octane, SGI) from a sagittal view.

Experiment 1

We presented knocking and drinking movements with 10 different affects and measured the ability of participants to categorise affect. We examined the perception of affect within the framework of a psychological space and related aspects of this psychological space to physical properties of the movements.

Methods

Movement Collection Knocking and drinking movements were recorded with affect. Two actors read a brief story that set the emotional scene for each movement. Measurements of the 10 affects (afraid, angry, excited, happy, neutral, relaxed, sad, strong, tired and weak) were obtained. This yielded a total of 120 movements (10 affects X 2 actors X 2 actions X 3 repetitions), however, due to recording difficulties data was lost for 2 movements of one actor.

Stimuli and Participants All 118 movements were displayed as above to fourteen Glasgow University student volunteers. Participants were naïve to the purpose of the study and were paid for their participation.

Design and Procedure Displays were blocked by the possible combinations of actor and action (2 actors x 2 actions). There were 4 trial blocks and a practice session of four trials. The order of blocks was randomised and participants were told that they would see a knocking or drinking arm movement. For each trial, participants viewed a computer display of the movement and were then presented with a dialog box that contained the names of the ten possible affects. Their task was to identify the affect by selecting one of the 10 choices.

Results

Over all the trials participants answered correctly 30% of the time; ranging from 15% (strong) to 50% (afraid) correct, this was significantly better than the chance value of 10% [t (13) = 20.3, $p < .005$, two-tailed]. Although the overall recognition rate was not high this

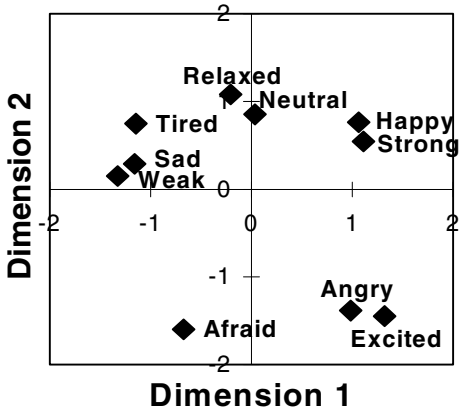


Figure 1. The psychological space obtained for Experiment 1

could be partially accounted for by some consistent misidentifications. For example, weak movements were identified as weak, sad or tired with equal frequency.

To better understand the structure of the results, a psychological space of the affects was constructed using the INDSCAL multidimensional scaling procedure (Kruskal and Wish, 1978). The 10x10 confusion matrix for each of the 4 block conditions was converted to measures of dissimilarity and input to the INDSCAL algorithm. The resultant solution was a unique 2-dimensional psychological space (see Figure 1) with r^2 = .87 and stress = .15. The first dimension accounted for approximately 70% of the variance and the second dimension for 17% of the variance. The two-dimensional structure of the psychological space is similar to that which would be predicted from a circumplex model of affect (see introduction) with the first dimension representing activation and the second dimension representing pleasantness.

Comparison of Psychological Space to Movement Kinematics.

We examined the movement kinematics to see whether any physical properties of the movement were related to either of the two dimensions defining the psychological space.

One of the striking things in the movement data is that the kinematic markers we measured consistently and smoothly differed between affects. For instance, sad movements were always slower than neutral movements and both these were slower than angry movement. This seems to correspond to the activation dimension from the models of affect. To test this, the kinematic markers were correlated to Dimension 1 and Dimension 2 co-ordinates of the 10 affects in the psychological space. Results of all these correlations are presented in Table 1, and Figure 2 shows an example of this relationship.

From Table 1 we can see the Dimension 1 (activation) co-ordinate of an affect correlated with the kinematic markers in such a way that energetic movements were positively correlated with shorter duration and greater magnitudes of average velocity, peak velocity, acceleration, deceleration and jerk. For Dimension 2 we found that, to a lesser extent there was a tendency of longer duration and smaller magnitude of the other kinematic markers to be correlated with positive affect. We examined this further by rotating the psychological space to find the orientation that maximised the r-squared values of the correlation with the six kinematic markers. It was found that a 27° counter-clockwise rotation resulted in the highest correlation with the kinematic markers with a r^2 value of 0.88 for Dimension 1 and 0.03 for Dimension 2. From these results it can be seen that while the original psychological space is roughly oriented so that energy in Dimension 1 is correlated with the speed of the movement, rotation of the space can improve the correlation.

Experiment 2

The aim of this experiment was to further investigate the role of speed in the recognition of affect from human movements. New sad, neutral and angry lifting and knocking movements were recorded from 3 women and 3 men. Through time warping (Bruderlin and Williams, 1995) the duration of movements were manipulated to change their speed. The original and morphed movements were displayed as point light stimuli to 10 participants who judged the intensity of affect.

Methods

Movement Collection Movements from 6 actors were recorded as they performed lifting and knocking actions with the three affects - angry, neutral and sad. As before, there were differences between kinematic markers for the movements. Angry movements had the

Table 1. Correlation of Movement Kinematics with Dimension 1 & 2 of psychological space. All values are Pearson's r and are significan at p<.oo5

Kinematic Properties	Dimension 1	Dimension 2
Duration	-0.85	0.65
Average Velocity	0.92	-0.49
Peak Velocity	0.91	-0.53
Peak Acceleration	0.83	-0.68
Peak Deceleration	-0.79	0.57
Jerk Index	0.83	-0.59

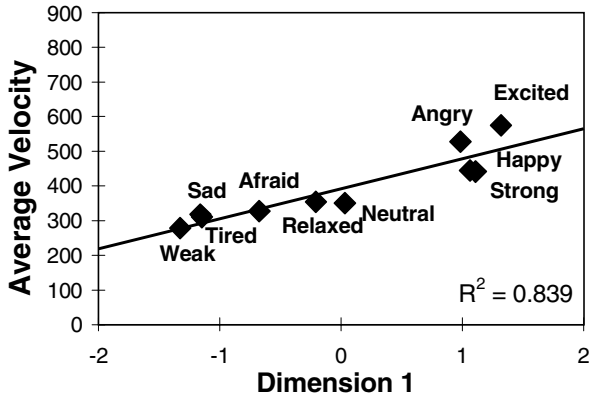


Figure 2. Plot of the average velocity of an affective movement versus the Dimension 1 coordinate obtained in the psychological space. Similar results are obtained for plots of the other kinematic markers versus Dimension 1.

shortest duration and highest velocities and sad the longest duration and lowest velocities.

For each movement the start and end points were defined using their velocity profile. Movement duration was defined as:

Duration = End point - Start point

For each actor a temporal step-size was calculated:

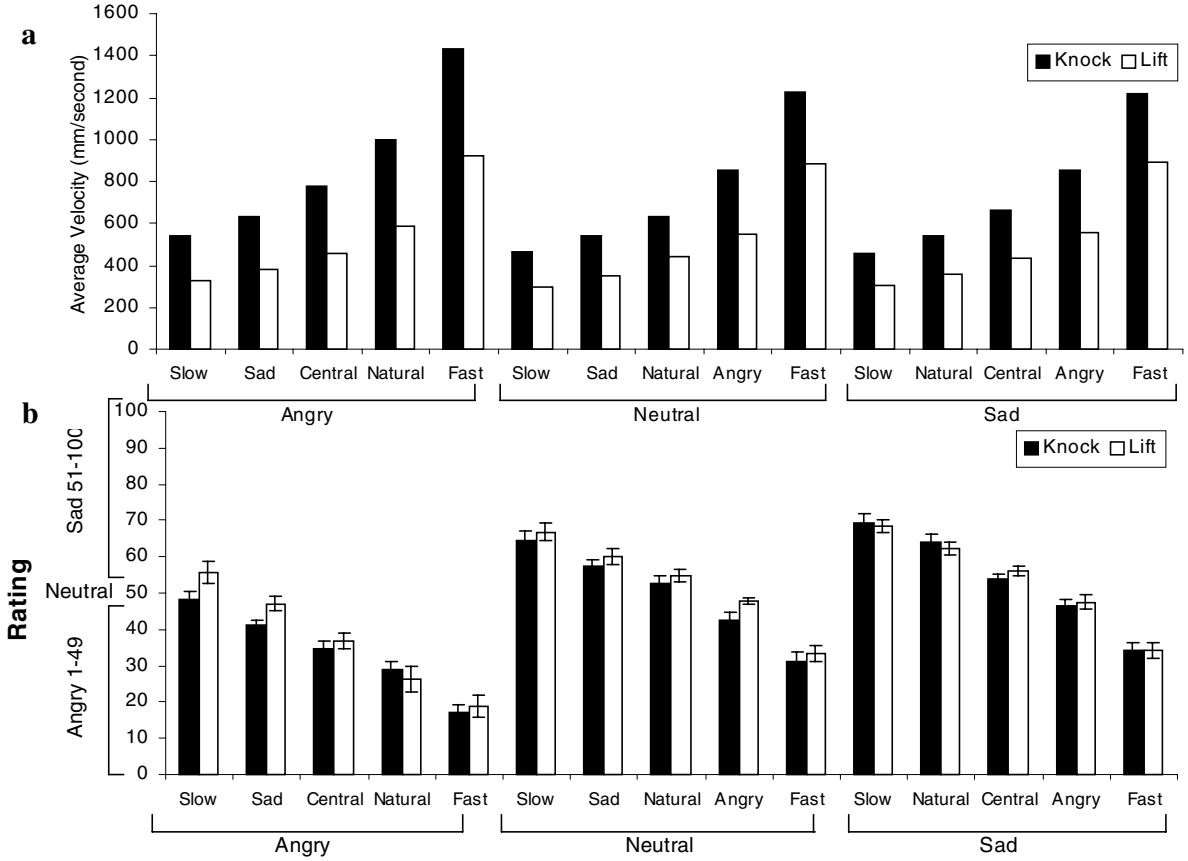
Step-size = (Sad Duration - Angry Duration)/2

The natural movements and the step-size calculation, were used in an interpolation algorithm (Hill and Pollick, 2000) to obtain 5 movements of differing duration for each affect. This yielded 15 movements for each actor, 3 of which were the natural movements. The interpolation preserved spatial properties of the movements such as the distance travelled by points between frames, but caused changes in spatio-tempotal properties, such as the average and peak velocities (see figure 3 and 4a). Movements had the following duration. A slow duration that was slower than the natural sad duration by one step-size; a natural sad duration; a central duration between angry and sad, this was very similar to the natural neutral movement duration; a natural angry duration; a fast duration, faster than the angry duration by 1 step-size. The central duration for neutral movements was the natural duration.

Participants Ten Glasgow University paid student volunteers participated in the experiment. All were naïve to the purpose of the study.

Design and Procedure Displays were blocked by action (2 actions). There were two blocks of trials, randomly ordered between participants. Participants were told that they would see human knocking and lifting movements. For each trial, participants viewed a computer display of the movement and were then presented with a dialog box that contained a 100-point scale for them to manipulate. First participants had to categorise the movement, then to rate its intensity.

A score between 1 and 49 indicates that the movement was perceived as angry, 1 being intense anger and 49, mild anger. A score of 51-100 indicates that the movement was perceived as sad, 51 being mildly sad and 100, intense sadness. A score of 50 indicates that the movement was neutral.



Frigure 3 a) The velocity of movements averaged across 6 actors and b) the rating participants gave to each movement, averaged across 6 actors and 10 participants. The x-axes illustrate the affect and duration used in the interpolation procedure. In b the rating scale is illustrated on the y-axis.

Results

For each affect there was a clear change in the classification and intensity ratings of affect as movement duration increased (figure 4b). This was the case for all three emotions, however, angry movements were seldom categorised as sad or neutral. These results are better illustrated when the rating data is correlated with kinematic markers, table 2 summarises these results.

Table 2. Correlation of kinematic markers with rating data, all values are Pearson's r and are significant at $p < .001$

Affect	Duration	Peak Velocity	Average Velocity
Angry	0.81	-0.67	-0.62
Neutral	0.91	-0.79	-0.79
Sad	0.96	-0.79	-0.82

Discussion

In this experiment the speeds of affective human arm movements were manipulated by changing movement duration. When new and original angry movements were viewed, they generally retained their identity as angry affect, but the intensity of perceived affect was modulated by velocity changes. Sad and neutral movements were similarly affected by changes in velocity, but faster movements were categorised as angry. These results further emphasise the role of velocity in affect, however, it is clear that there are other properties of the movements that were not controlled by velocity, but that play a role in the discrimination of affect. Currently it is only possible to speculate about what these properties are, but they may include spatial relationships, such as posture, between points of the displays; phase relations between the points or more dynamic properties, such as perceived force. It is clear, however, that velocity plays a major role in the recognition of affect from biological motion displays, particularly in modulating the intensity of perceived affect.

General Discussion

The results of Experiment 1 showed that the perceived affect of arm movements conformed well to models of experienced affect (Russell, 1980; Watson and Tellegen, 1985; Thayer, 1989, Larsen and Diener; 1992; Feldman Barrett and Russell, 1999). Moreover, the activation axis of these models was correlated to physical characteristics of the movement in a consistent manner such that greater activation was related to greater magnitudes of velocity, acceleration and jerk of the movement.

The finding of a circumplex structure for perceived affect is consistent both with duality between the perception and production of movement as well as a role for high-level information in the interpretation of motion derived from human movement. In addition, the continuous structure of the circumplex model parallels the smoothly varying range of speeds with which a movement can be performed. Thus, it would appear that the mapping between stimulus properties and representation of affect is a fairly direct one for the activation axis. However, such a direct connection between stimulus and representation has proven elusive for the second dimension of pleasantness. Other research has suggested that subtle phase relations between the joints (Amaya, Bruderlin and Calvert, 1996) might possibly carry information about affect.

The results from Experiment 2 further showed the way in which speed modulates interpretation of affective movements. However, factors controlling dimension 2 of the psychological space, could not be entirely discounted. For sad and neutral movements it was possible to change the percept to be angry, but angry movements remained discriminable as angry at most speeds. This suggests that angry movements are distinct form the other movements in some other way and that humans are sensitive to this difference. Indeed, angry, afraid and excited movements fell at a different location on the second dimension in Experiment 1, than the other movements. So perhaps whatever properties in the movements control this dimension also act as "danger indicators". From an evolutionary point of view such indicators makes good sense and it could be argued that this would enhance the models of experienced affect, since pleasantness may be just another way of discriminating emotions that a person seek out, from those they avoid.

References

Amaya, K., Bruderlin, A. & Calvert, T. (1996). Emotion from motion. In *Graphics Interface '96*, W.A. Davis & R. Bartels, Eds., pp 222-229.

Barclay, C.D., Cutting, J.E. & Kozlowski, L.T. (1978). Temporal and spatial factors in gait perception that influence gender recognition. *Perception and Psychophysics, 23*, 145-152.

Bertenthal, B.I. & Pinto, J. (1994). Global processing of biological motions. *Psychological Science, 5*, 221-225.

Bruderlin, A. & Williams, L. (1995). Motion Signal Processing. In *Graphics Interface '95*, W.A. Davis & R. Bartels, Eds., pp 97-104.

Cutting, J.E. & Kozlowski, L.T. (1977). Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the Psychonomic Society, 9*, 353-356.

Cutting, J.E. (1978). Generation of synthetic male and female walkers through manipulation of a biomechanical invariant. *Perception, 7*, 393-405.

Cutting, J.E., Proffitt, D.R. & Kozlowski, L.T. (1978). A biomechanical invariant for gait perception. *Journal of Experimental Psychology: Human Perception and Performance, 4*, 357-372.

Decety, J. & Grezes, J. (1999). Neural mechanisms subserving the perception of human actions. *Trends in Cognitive Sciences, 3*, 172-178.

Dittrich, W.H., Troscianko, T., Lea, S.E.G. & Morgan, D. (1996). Perception of emotion from dynamic point-light displays represented in dance. *Perception, 25*, 727-738.

Flash, T & Hogan, N. (1985). The coordination of arm movements: An experimentally confirmed mathematical model. *Journal of Neuroscience, 5*, 1688-1703.

Hill, H & Pollick, F.E. (in press). Exaggerating temporal differences enhances recognition of individuals from point light displays. *Psychological Science*.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics, 14*, 201-211.

Kozlowski, L.T. & Cutting, J.E. (1977). Recognizing the sex of a walker from a dynamic point-light display. *Perception and Psychophysics, 21*, 575-580.

Kruskal, J.B. & Wish, M. (1978). *Multidimensional Scaling*, Sage University Paper series on Quantitative Applications in the Social Sciences, 07-011, Beverly Hills and London: Sage Publications.

Mather, G., Radford, K., & West, S. (1992). Low-level visual processing of biological motion. *Proceedings of the Royal Society of Londdon B, 249*, 149-155.

Mather, B. & Murdoch, L. (1994). Gender discrimination in biological motion displays based on dynamic cues. *Proceedings of the Royal Society of Londdon B, 258*, 273-279.

Neri, P., Morrone, M., & Burr, D. (1998). Seeing biological motion. *Nature, 395*, 894-896.

Shiffrar, M. & Freyd, J.J. (1990). Apparent motion of the human body. *Psychological Science, 1*, 257-264.

Shiffrar, M. & Freyd, J.J. (1993). Timing and apparent motion path choice with human body photographs. *Psychological Science, 4*, 379-384.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3*, 131-141.

Runeson, S (1994). Perception of biological motion: The KSD-principle. In *Perceiving Events and Objects*, G. Jansson, S.S. Bergstrom & W. Epstein, Eds., pp 383-405.

Runeson, S., & Frykholm, G. (1981). Visual perception of lifted weight. Journal of Experimental Psychology: Human Perception and Performance, 7, 733-740.

Runeson, S., & Frykholm, G. (1983). Kinematic specification of dynamics as an informational basis for person and action perception: Expectation, gender recognition, and deceptive intention. *Journal of Experimental Psychology: General, 112*, 585-615.

Thornton, I.M., Pinto, J. & Shiffrar, M. (1998). The visual perception of human locomotion. *Cognitive Neuropsychology, 15*, 535-552.

Walk, R.D. & Homan, C.P. (1984). Emotion and dance in dynamic light displays. *Bulletin of the Psychonomic Society, 22*, 437-440.

Graph-based Reasoning: From Task Analysis to Cognitive Explanation

David Peebles (djp@psychology.nottingham.ac.uk)
Peter C.-H. Cheng (pcc@psychology.nottingham.ac.uk)
ESRC Centre for Research in Development, Instruction and Training,
Department of Psychology, University of Nottingham, Nottingham, NG7 2RD, U.K.

Abstract

Models of graph-based reasoning have typically accounted for the variation in problem solving performance with different graph types in terms of a task analysis of the problem relative to the particular visual properties of each graph type (e.g. Lohse, 1993; Peebles, Cheng & Shadbolt 1999, submitted). This approach has been used to explain response time and accuracy differences in experimental situations where data are averaged over experimental conditions. A recent experiment is reported in which participants' eye movements were recorded while they were solving various problems with different graph types. The eye movement data revealed fine grained scanning and fixation patterns that are not predicted by standard task analytic models. From these eye-movement studies it is argued that there is a missing level of detail in current task analytic models of graph-based reasoning.

Introduction

The ability to retrieve and reason about information in graphs and diagrams is a skill which requires the complex interaction of three primary elements: the cognitive abilities of the user, the graphical properties of the external representation, and the requirements of the task. Several frameworks have been proposed to understand interactive behaviour of this sort. In the area of graph-based reasoning, Peebles, Cheng & Shadbolt (1999, submitted) have proposed the GBR model incorporating these three factors. Gray (2000; Gray & Altmann, 2000) has proposed the *Cognition-Task-Artifact triad* within which to characterise interactive behaviour in the related context of human-computer interaction. This latter framework has recently been further developed by Byrne (in press) to encompass the perceptual and motor capabilities of the user, termed *Embodied Cognition*.

The main aim of these models and frameworks is to aid the development of detailed cognitive models of the cognitive, perceptual and motor processes involved in the tasks under study. Constructing cognitive process models that are grounded in cognitive theory allows the incorporation and testing of relevant cognitive factors such as the required declarative and procedural knowledge, the strategies adopted, and the limitations of working memory. This approach contrasts with that of *cognitive task analysis* which simply specifies the cognitive steps required to perform the task.

In the area of graph-based reasoning, Lohse (1993) developed the GOMS class of task analysis techniques (Card, Moran, & Newell, 1983; Olson & Olson, 1990; John & Kieras, 1994) by including additional cognitive parameters to produce a cognitive model which simulates how people answer certain questions using line graphs, bar graphs and tables. Lohse's model was based on the assumption that graph knowledge is represented as graph schemas (Pinker, 1990) which allow the recognition and interpretation of different classes of graph. Included in a graph schema are task-specific rules that define sequences of procedures for retrieving information from the graph given a particular information-retrieval task. Lohse's model predicted the time to answer a given question by assuming that people scanned the graphical representation in a manner which produced an optimal sequence of eye movements that minimized the number of saccades and fixations to reach the target location.

In the *Graph Based Reasoning* (GBR) model (Peebles et al., 1999, submitted), a similar set of assumptions was employed to explain several results of experiments investigating the factors affecting reasoning with *informationally equivalent* (Larkin & Simon, 1987) graphs of different types from the same general class; Cartesian coordinate (x–y) graphs. Figure 1 shows the types of graph used in our experiments. The graphs are informationally equivalent as the both encode the same two functions between time and the variables A and B. The *Function* graph in Figure 1a represents time on the *x* axis and the A and B variables on the *y* axis whereas the *Parametric* graph in Figure 1b represents the A and B variables on the *x* and *y* axes respectively while time is plotted as a parameterizing variable along the curve.

Although the two graphs assign different variables to their axes, they would be considered similar in several important ways identified in the literature. Firstly, both are *Cartesian* graphs using a two dimensional coordinate system to relate quantities and represent magnitudes. It is likely, therefore, that both graphs invoke similar general schemas and interpretive processes (Pinker, 1990; Kosslyn, 1989). Secondly, both are simple line graphs and consequently share many of the same general interpretive rules. Furthermore, it is likely that inferences from both graphs are influenced by the same set of biases (Carpenter & Shah, 1998; Gattis & Holyoak, 1996; Shah & Carpenter, 1995). Finally, the graphs are infor-

mationally equivalent as they have been generated from the same data set.

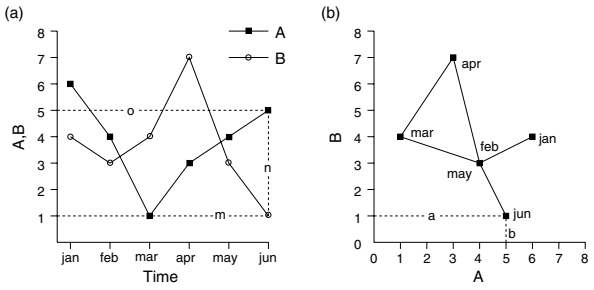


Figure 1: Informationally equivalent function and parametric graphs

Despite these similarities, however, in previous experiments we have demonstrated that for a wide range of questions, parametric and function graph users differ substantially in both the time it takes to respond and in their rates and patterns of errors (Peebles et al., 1999, submitted). The GBR model has been successful in explaining why such differences occur with these graph types despite their many common properties. Using the graphs in Figure 1 as an example, we found in our experiments that when participants were asked to retrieve the value of A when the value of B is 1, responses from parametric graph users were significantly more rapid and accurate than those from the function graph users. The GBR model explains these differences in terms of the optimal visual scan path the users follow through the graph. The variability in responses is apparent from the sequence of hypothesised saccades in the two graphs. In Figure 1a, the sequence of saccades is *m*, *n*, *o*, whereas in Figure 1b the process requires just two saccades, as shown by the line sequence *a*, *b*. The higher probability of an erroneous response using the function graph was explained by the additional number of possible incorrect saccades that the function graph users may make.

Although these optimality assumptions are useful in that they provide an account of differences in mean RT and error data for the different graph conditions, it remains an open question, however, whether they gloss over important cognitive and strategic factors at an individual level. For example, graph users may be required to re-encode items of information that have been lost from working memory during the course of processing. In addition, given that graph users are aware that information is available for re-scanning at all times, it is possible that they may make a strategic decision to trade off additional saccades for a reduction in working memory load. If this is the case, then the current analyses may miss out an important level of detail which sheds light on the cognitive load that these tasks are imposing and the strategies by which graph users optimise their retrieval procedures. Furthermore, information at this level of detail will provide valuable constraints on cognitive models of these reasoning processes.

To address these issues, we devised an experiment in which participants were asked to solve some simple tasks using different graph types of the same general class which, based on the optimality assumptions above, would be predicted to produce different response patterns. These predictions can be elaborated in terms of an optimal sequence of fixations required to solve the given task. To test these optimality assumptions and predictions, therefore, some of the participants' eye movements would be recorded as they solved the problems.

One of the most common tasks carried out when using a graph is to elicit the value of one variable corresponding to a given value of another. This task was chosen for the experiment as it is so widely performed and because the procedures involved are relatively simple. The knowledge required to carry out these tasks is primarily the sequence of fixations required to reach the *given location* in the graph representing the given value of the given variable and then from there to the *target location* representing the corresponding value of the required variable. In previous research, however, we have discovered that the effectiveness of a particular graphical representation for retrieving the required information depends on the details of the task, i.e. which variable is given and which is sought (Peebles et al., 1999, submitted).

Experiment

Method

Participants and materials Forty-four undergraduate and postgraduate psychology students from the University of Nottingham were paid £3 to take part in the experiment. The experiment was carried out using two PC computers with 17 in displays. A further four participants from the same population were paid £5 to participate in the eye-movement study. The eye tracker employed in the experiment was an SMI iView system using a RED II desktop pupil/corneal reflectance tracker with a sampling rate of 50 Hz. This system records eye movements at 20 ms intervals remotely from a position in front of the experimental computer display. Although the system contains an automatic head movement compensation mechanism, to further reduce recording error due to head movement, participant's heads were restrained in a frame fixed to the table.

The stimuli used in the experiment were four graphs, shown in Figure 2, depicting the amount (in millions of units) of UK offshore oil and gas production between two decades, 1970–1979 and 1980–1989. The graphs and data sets were designed so that the independent variable (IV—year) and the two dependent variables (DVs—oil and gas) all had ten values ranging from 0 to 9 and that the full range of these values was represented by the data points for oil and gas in both decades.

Participants were seated approximately 80 cm from the 72 ppi computer display. The graphs were 15.5 cm square (including axis labels), corresponding to approximately 11.1° of visual angle. The characters representing variable values were 0.4 cm high (approximately .21° of

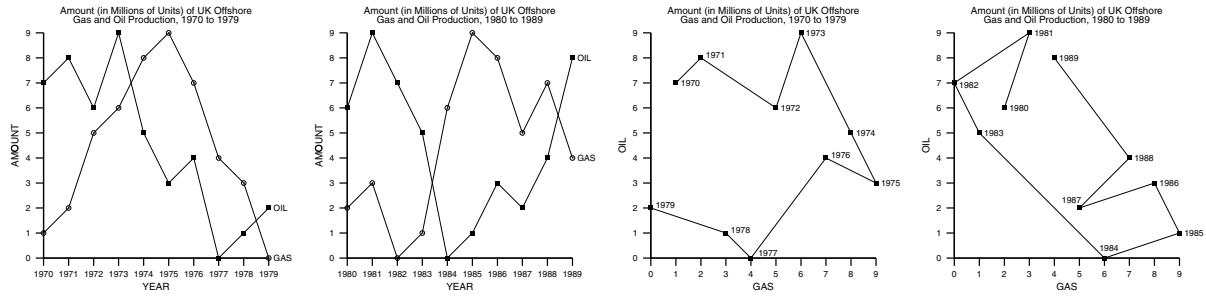


Figure 2: Function and Parametric Graphs Used in the Experiment

visual angle) while those for the axis labels and questions were 0.4 cm and 0.5 cm high (approximately .29° and .36° of visual angle) respectively. Axis ticks were spaced 1.5 cm (approximately 1.1° of visual angle) apart.

The full range of values for each of the variables was used to produce 120 questions. These questions all had the same basic structure and were of three types; DV–DV and DV–IV questions gave the value of one of the dependent variables and required the corresponding value of the second DV or the IV respectively, while IV–DV questions gave a value of the independent variable and required the corresponding DV value to be produced. There were 20 of each of question type and participants were required to answer all 60 for both decade graphs, producing a total of 120 questions.

Design and Procedure The experiment was a mixed design with one between-subjects variable, (graph type) and two within-subjects variables (question type and graph number). Participants were randomly allocated to one of the two graph type conditions producing a total of 22 participants per condition in the main experiment and two participants per condition in the eye movement study. During the experiment, the two graphs were presented alternately with the first graph being selected at random. On each trial, a graph would be presented with a question above it. The questions were presented in a form so that the minimum amount of text was shown. For example, the question GAS = 2, OIL = ? requires the value of oil when gas is equal to 2 to be found. When a year value was required, the final items of text in the question would be YEAR = 197? or YEAR = 198? depending on the current graph being presented and participants were instructed beforehand to enter only the final number of the target year. Each element of the question was centered on a co-ordinate point which remained invariant throughout the experiment with approximately 3.5 cm (approximately 2.5° of visual angle) between the centres of adjacent text items. Together with the graph and question, a button labelled *Answer* appeared in the top right corner of the window. Participants were instructed to click on this answer button as soon as they had obtained the answer to the question. Response times were recorded from the onset of a question to the mouse click on the answer button. When this button was clicked upon, the button, graph and question were removed from the screen and a circle of buttons labelled clockwise from 0 to 9 appeared centered on the answer button. Participants entered their answers by clicking the appropriate number button. When the number button was clicked, the next graph, question, and answer button appeared on the screen. This method was devised so that participants in the eye movement study would not have to take their eyes away from the screen to enter answers, as would be the case if using the keyboard.

Before starting the experiment, participants were given as much time as necessary to become familiar with the two graphs in their condition and were also provided with an opportunity to practice entering numbers using the circle of number buttons and the mouse. Participants were asked to answer the questions as rapidly and as accurately as possible

Results

Response accuracy and latency data The proportions of correct responses and mean response times (RTs) for each of the question types for the two graphs in each condition are presented in Figure 3. Confirming the relative simplicity of the experimental tasks, the data reveal high levels of accuracy for all three question types in both graph conditions. An ANOVA on the response accuracy data, however, revealed a significant effect of question type $F(2, 239) = 28.187$, $p < 0.01$, $MSE = 0.123$ indicating that some types of question were generally more demanding than others. The nature of this effect can be clearly seen in Figure 3. In both graph conditions, more errors were made carrying out the DV–DV task than the other two while the IV–DV task was the most accurately responded to.

While there is little variability in the accuracy of responses between conditions, the time taken by participants in the two groups to make these responses varies significantly both between conditions and within each condition according to the type of question being attempted. An ANOVA on the RT data revealed significant effects of question type $F(2, 239) = 18.447$, $p < 0.01$, $MSE = 4974038$, and graph number $F(1, 239) = 5.76$, $p < 0.05$, $MSE = 1223302$ and significant interactions be-

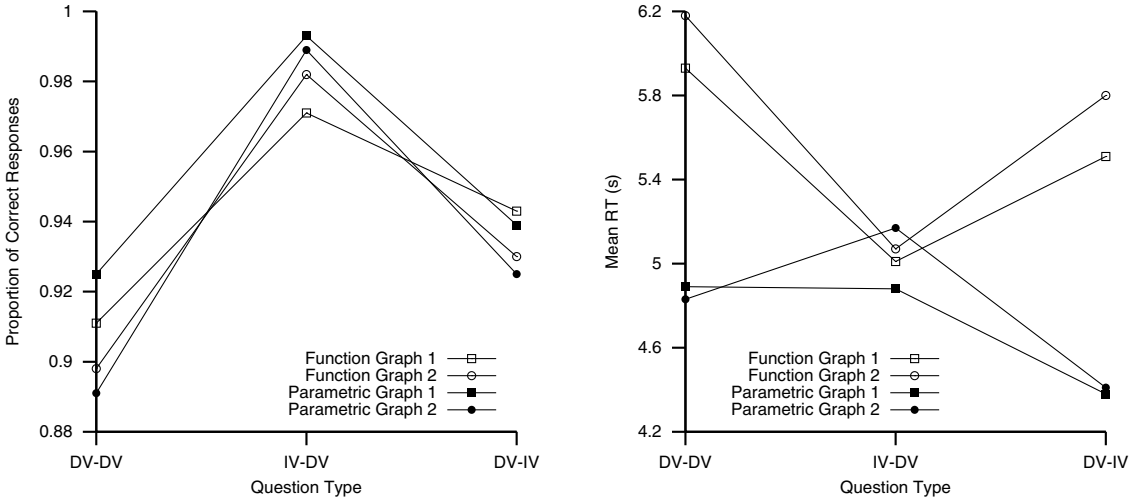


Figure 3: Plots of mean correct responses and RTs for function and parametric graph conditions for each question type

tween graph type and question type $F(2, 239) = 36.314$, $p < 0.01$, $MSE = 9791754$ and between graph type, question type and graph number $F(2, 239) = 3.913$, $p < 0.05$, $MSE = 466423$. The nature of these effects and complex interactions is apparent in Figure 3. In both conditions, it takes approximately 5 s to read the question and retrieve the required DV value for a given year. However, to carry out the reverse task and find the year corresponding to a given DV value takes, on average, over 1 s longer when using the function graph than when using the parametric graph. A similar disparity in RT is found when the task is to retrieve a DV value corresponding to a given DV value.

In both conditions, errors are evenly distributed over experiment trials. The mean proportion of correct responses over the first 10 trials for function and parametric graphs is .91 and .94 respectively. Over the course of the experiment, the mean RT for both conditions reduced by approximately 2 s, the rates of these reductions being described by power functions with similar slopes.

To analyse the results of the experiment, the display was divided into five regions in a manner similar to that employed by Carpenter and Shah (1998). The regions, shown in Figure 4, were the same for all four graphs and define the relevant units of the display for the fixation analysis: *question, graph pattern, x-axis, y-axis*, and *answer* buttons.

The pattern of RT data from the experiment can be explained by the GBR model using the optimality assumptions and fixation predictions outlined above. The significant increase in time to answer DV–IV questions using the function graphs is due to the fact that in the parametric graphs, the target values are positioned next to the given location so that the additional cognitive and perceptual processes required to fixate on the target location are not required. In this case the optimal sequence of fixations is predicted to be: *question, axis, graph, answer* whereas that for the function graphs is: *question, axis, graph, axis, answer*.

The DV–DV questions are of the same type as the example question given in the introduction and so the smaller mean RT in the parametric condition can be accounted for in terms of the previous explanation, namely, that to reach the target location in the function graphs requires an additional saccade and fixation and the associated cognitive operation to retrieve a further step in the process. So, the optimal sequence of fixations for parametric graphs is predicted to be: *question, axis, graph, axis, answer*, whereas that for the function graphs is: *question, axis, graph, graph, axis, answer*.

For the IV–DV questions, the relative rapidity with which function graph users are able to answer these questions compared to others is due to the fact that they are able to rapidly identify the given year on the x axis and then carry out the two step process of identifying the target point on the correct line and retrieving its value from the y axis. The optimal sequence of fixations for this procedure is: *question, axis, graph, axis, answer*. The data show that this procedure takes approximately the same time as the corresponding procedure for the parametric graphs which requires the search of the given year in the graph and the retrieval of its value from the target axis, the optimal fixation sequence of this procedure being: *question, graph, axis, answer*.

The results of the main experiment show that, despite the numerous similarities that exist between function and parametric graphs, the type of graph used can significantly affect the time it takes to retrieve the required information and that this effect is dependent on the nature of task. The experiment also showed that the probability of retrieving incorrect information depends on specific details of the task, i.e. which variable is given in the ques-

tion and which variable value is being sought. The GBR model explains these differences in terms of a detailed task analysis and the assumption of an optimal scan path through the graph to the target location.

Eye movement data To analyse the eye movement data, the raw x and y co-ordinate data from the eye tracker were aggregated into *gazes*—sequences of consecutive fixations on a display region unbroken by fixations in other regions (Carpenter and Shah, 1998). The minimum duration of a gaze was defined as 100 ms as this value was sufficiently large to eliminate most saccades, short fixations and noise in the data while still capturing all the relevant fixations. The data from each participant were analysed so that gazes of 100 ms or more in each region were recorded and a scan path consisting of the sequence of gazes for each question was produced.

Several interesting patterns emerge from the analysis of these gaze sequences. Firstly, the average number of transitions between regions for all questions types, shown in Table 1, is consistently greater than the optimal number predicted by the GBR model. For all of the question types, and irrespective of the type of graph being used, participants made, on average, between three and four additional transitions in order to reach the solution. In the majority of cases, these additional transitions were between the axes and the graph and the question and the graph as participants rarely fixated upon the answer region until entering an answer. In 31% of all trials, participants made at least one additional gaze on an axis after having previously fixated upon that axis and then the graph. A detailed visual analysis of the raw eye movement data for these trials revealed that in most cases, participants had fixated upon a given axis value and then proceeded to the plot point in the graph corresponding to that value. Upon reaching this point, an additional saccade was then made to the axis to check that the value was in line with the point.

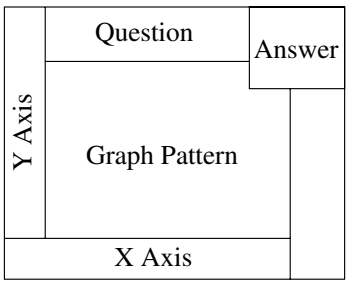


Figure 4: Five regions of the display defined for the fixation analysis

From the eye movement data analysis, it is clear that, although the participants did, in general, solve the various problems by following the optimal gaze paths characterised by the GBR model, they made considerably more gazes than is predicted by the model. Although it is likely that many of these additional transitions are due to checking procedures of the sort outlined above, it is possible that common patterns in the gaze sequences indicate limitations of working memory or problem solving strategies adopted by graph users. For example, in 62.7% of all trials and irrespective of the question type being attempted, participants made at least one additional gaze on the question after having initially gazed upon the question and subsequently the graph. This pattern suggests two possible explanations. The first is that participants have initially encoded the three elements of the question but are required to re-encode certain parts of it that are unable to be retrieved from working memory due to the cognitive load involved in carrying out the problem solving procedures. The second explanation is that participants have adopted a strategy by which only the initial part of the question is encoded and the second part is encoded only when required. According to this explanation, in the majority of trials, participants effectively break the problem into two sections, the first to get to the given location in the graph, the second to move from the given location to the target location corresponding to the solution. It is also possible that the observed gaze patterns may result from a combination of these factors if, during the course of the experiment, participants adopt the above strategy in order to minimise the number of question element retrieval failures.

Table 1: Mean number of gaze transitions between display regions for Function and Parametric graphs observed (Obs) for each question type, and the optimal (Opt) number predicted by the GBR model

Question	Function		Parametric	
Type	Obs	Opt	Obs	Opt
DV–DV	7.66	5.0	8.21	5.0
IV–DV	7.86	5.0	8.90	4.0
DV–IV	8.05	5.0	8.05	4.0

Discussion

Reasoning with Cartesian graphs involves a complex interaction between the perceptual and cognitive abilities of the reasoner, the visual properties of the graph, and the specific task requirements. Models of graph-based reasoning (e.g. Lohse, 1993; Peebles et al., 1999, submitted) have largely focussed on providing a detailed analysis of the task in relation to the the visual properties of the graph and explaining differences in performance in terms of the interaction of these two elements. These models have been successful in accounting for variations in aggregate RT data between users of different graph types by characterising an optimal sequence of fixations based on the task analysis that will achieve the goal. Error data is also explained by hypothesising sets of plausible deviations from these optimal sequences.

To produce detailed cognitive models of graph use grounded in cognitive theory, however, then the third,

cognitive element of the triad must be fully incorporated into these accounts. The explanatory and predictive power of cognitive models in complex interactive domains compared to cognitive task analyses has been demonstrated (e.g. Gray, John, & Atwood, 1993). By incorporating such cognitive factors as the user's knowledge, strategies and working memory capacity into graph-based reasoning models, the explanatory and predictive power of these models can be increased and greater insights into the processes and factors affecting these complex interactions can be obtained.

Although the standard experimental variables of RT and error rates provide some information upon which to formulate and test cognitive hypotheses, much richer data is obtained when eye movements are recorded during the experiment. In such a visual domain as graph-based reasoning, eye movements are an important source of information regarding how people acquire and process graphical information and the strategies they adopt when interpreting and working with graphs. This has been demonstrated by Carpenter and Shah (1998) in their analysis of eye movements in graph comprehension tasks which revealed the cyclic nature of the pattern recognition and cognitive processes involved in graph comprehension.

In contrast, the present experiment provides an example of how eye movement data can be used in the analysis of more goal directed graph-based reasoning tasks in which the aim of the interaction is not to simply understand the graph but to retrieve specific information from it. The results of the main experiment showed that the ability of people to retrieve the same information from computationally inequivalent but visually similar Cartesian graphs can be significantly affected by the type of graph used. A plausible explanation of these differences can be provided by the GBR model in terms of an analysis of the task and an assumption of the optimal scan path through the graph to the target location representing the problem solution. These results support and extend the findings of previous experiments (Peebles et al., 1999, submitted) and provide further evidence that the GBR model can account for data that cannot be explained solely in terms of the visual properties of the graphs.

The actual scan paths revealed by the eye movement study show, however, that these optimality assumptions serve as an approximation that can be applied to data aggregated over experimental conditions but which tend to obscure the detailed sequences of saccades made by individuals. It is clear that further research is required to investigate the cognitive factors underlying these saccade patterns in greater detail. It is also clear, however, that cognitive models of graph-based reasoning must incorporate more sophisticated cognitive mechanisms in order to account for these findings.

Acknowledgements

This research is funded by the UK Economic and Social Research Council through the Centre for Research in Development, Instruction and Training.

References

Byrne, M. D., (in press). ACT-R/PM and menu selection: Applying a cognitive architecture to HCI. *International Journal of Human-Computer Studies*.

Card, S. K., Moran, T. P., & Newell, A., (1983). *The psychology of human-computer interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, *4*, 75–100.

Gattis, M., & Holyoak, K. (1996). Mapping conceptual to spatial relations in visual reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 231–239,

Gray, W. D. (2000). The nature and processing of errors in interactive behaviour. *Cognitive Science*, *11*, 205–248.

Gray, W. D., & Altmann, E. M. (2000). Cognitive modeling and human-computer interaction. In W. Karwowski, (Ed.), *International encyclopedia of ergonomics and human factors*. New York: Taylor & Francis, Ltd.

Gray, W. D., John, B. E., & Atwood, M. E. (1993). Project Ernestine: Validating a GOMS analysis for predicting and explaining real-world performance. *Human-Computer Interaction*, *8*, 237–309.

John, B. E., & Kieras, D. E., (1994). *The GOMS family of analysis techniques: Tools for design and evaluation*. (Tech. Rep. CMU-HCII-94-106). Pittsburgh, PA: Carnegie Mellon University, Human-Computer Interaction Institute.

Kosslyn, S. M., (1989). Understanding charts and graphs. *Applied Cognitive Psychology*, *3*, 185–226

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, *11*, 65–100.

Lohse, G. L. (1993). A cognitive model for understanding graphical perception. *Human-Computer Interaction*, *8*, 353–388.

Olson, J. R., & Olson, G. M., (1990). The growth of cognitive modeling in human-computer interaction since GOMS. *Human-Computer Interaction*, *5*, 221–265.

Peebles, D., Cheng, P. C.-H., & Shadbolt, N. (1999). Multiple processes in graph-based reasoning. In *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.

Peebles, D., Cheng, P. C.-H., & Shadbolt, N. (submitted). A model of graph-based reasoning: Integrating the role of visual features, knowledge and search.

Shah, P., & Carpenter, P. A. (1995). Conceptual limitations in comprehending line graphs. *Journal of Experimental Psychology: General*, *124*, 43–62.

The Impact of Feedback Semantics in Visual Word Recognition: Number of Features Effects in Lexical Decision and Naming Tasks

Penny M. Pexman (pexman@ucalgary.ca)
Department of Psychology, University of Calgary
2500 University Drive NW, Calgary AB, T2N 1N4

Stephen J. Lupker (lupker@julian.uwo.ca)
Department of Psychology, The University of Western Ontario
London, ON, Canada, N6A 5C2

Yasushi Hino (hino@sccs.chukyo-u.ac.jp)
Department of Psychology, Chukyo University
101-2 Yagotohonmachi, Showaku, Nagoya, Aichi 466-8666, Japan

Abstract

The notion of feedback activation from semantics to both orthography and phonology has recently been used to explain certain semantic effects in visual word recognition, including polysemy effects (Hino & Lupker, 1996; Pexman & Lupker, 1999) and synonym effects (Pecher, in press). In the present research we tested an account based on feedback activation by investigating a new semantic effect: number of features (NOF). Words with high NOF (e.g., LION) should activate richer semantic representations than words with low NOF (e.g., LIME). Richer semantic representations should facilitate lexical decision task (LDT) and naming task performance via feedback activation to orthographic and phonological representations. The predicted facilitory NOF effects were observed in both LDT and naming.

Introduction

Although the average speaker or reader of English seldom notices it, the English language is actually quite ambiguous in its usage. For example, many English words are "polysemous", in that they have multiple meanings (e.g., BANK). Thus, deriving the intended meaning requires the use of context. These polysemous words have been a useful tool in psycholinguistic research since they allow researchers the opportunity to study the impact of semantic ambiguity on word recognition and reading.

There is now considerable evidence that semantic ambiguity produces a processing advantage in lexical decision tasks (LDT) and naming tasks. That is, responding in those tasks is usually faster to polysemous than nonpolysemous words (Borowsky & Masson, 1996; Gottlob, Goldinger, Stone, & Van Orden, 1999; Hino & Lupker, 1996; Hino, Lupker, Sears, & Ogawa, 1998; Jastrzembski, 1981; Jastrzembski & Stanners, 1975; Kellas, Ferraro, & Simpson, 1988; Lichacz, Herdman, LeFevre, & Baird, 1999; Millis & Button, 1989; Pexman & Lupker, 1999; Rubenstein, Garfield, & Millikan, 1970). This effect has proven difficult to explain for current models of word recognition. For example, Joordens and Besner (1994) attempted to simulate polysemy effects using two PDP models but found that neither model was successful. The problem is that polysemy involves a one-to-many mapping between orthography and semantics and, thus, polysemous words should create competition in the semantic units. Because Joordens and Besner assumed that lexical decision performance depends on the settling time in the semantic units, the inevitable result was that this competition hindered, rather than facilitated, performance. That is, according to this and similar models, polysemy should produce a processing disadvantage in LDTs (for related discussions see Besner & Joordens, 1995; Kawamoto, Farrar, & Kello, 1994; Borowsky & Masson, 1996; Piercey & Joordens, 2000; Rueckl, 1995).

As Hino and Lupker (1996) argued, however, it is possible to explain polysemy effects within a PDP framework if slightly different assumptions are made. Following Balota, Ferraro, and Connor's (1991) basic argument, Hino and Lupker assumed that semantic activation feeds back to the orthographic units. That is, when a target word is presented, there is initially activation of an orthographic representation for that word. Very quickly, there is also activation of a semantic representation for the target word (and also activation of a phonological representation). The semantic representation then increases the activation of the orthographic (and phonological) representation via feedback connections. Because polysemous words (e.g., BANK) have a more extensive semantic representation than nonpolysemous words, polysemous words would produce more semantic activation than nonpolysemous words. Hence, the feedback activation from semantics to orthography should be stronger for polysemous words than for nonpolysemous words. As a result, the activation in the orthographic units for polysemous words should increase more rapidly than that for nonpolysemous words. Assuming that lexical decision responses are mainly based on orthographic activation, the expectation is that LDT responses should be faster for

polysemous than for nonpolysemous words, as is typically observed.

The explanation for polysemy effects in naming tasks is similar. For polysemous words, there would be considerable semantic activation, which would then help activate the phonological (as well as the orthographic) units. This semantic activation of phonological units could happen two ways: via feedforward connections for orthography-semantics-phonology linkages, and also via feedback connections for orthography-phonology-semantics-phonology linkages. In a naming task, it is assumed that responses are based on activation in the phonological units. Polysemous words would receive more phonological activation (via semantics), which would lead to a processing advantage in the naming task. Thus, according to Hino and Lupker (1996), polysemy effects in both tasks can be readily explained within a fully-interactive, PDP-type model of word recognition if feedback activation is assumed to play an important role in the process.

Note that certain models of word recognition do assume an important role for feedback connections. For example, Van Orden and Goldinger (1994; see also Stone, Vanhoy & Van Orden, 1997) argued for a system that incorporated both feedforward and feedback activation between sets of processing units. Additionally, in Seidenberg and McClelland's (1989) PDP model, feedback connections were proposed, although they were never implemented. Feedback connections from semantic to orthographic units were also included in some of Plaut and Shallice's (1993) simulations. Thus, models of this sort would be quite consistent with the existence of polysemy effects.

What should also be noted is that polysemy effects are not the only effects in the word recognition literature consistent with Hino and Lupker's (1996) feedback activation account. For example, Pexman, Lupker, and Jared (2001) argued that a feedback activation explanation, involving feedback from the phonological to the orthographic units, was required in order to explain homophone effects. Homophones are words like MAID and MADE for which multiple spellings (and meanings) correspond to a single phonological representation. As had been typically reported (e.g., Rubenstein, Lewis, & Rubenstein, 1971), homophones produced longer lexical decision response latencies than control words in Pexman et al.'s experiments. These homophone effects were most apparent for low frequency homophones with high frequency homophone mates, and were larger in LDT when pseudohomophones (e.g., BRANE) were used as foils (as compared to pseudoword foils, e.g., PRANE).

In terms of the feedback activation account, homophone effects are assumed to be caused by a single phonological representation activating two orthographic representations (Pexman et al., 2001) while polysemy effects are presumed to be caused by multiple semantic representations activating a single orthographic representation (Hino & Lupker, 1996). That is, in spite of the fact that these two effects go in opposite directions, they are both presumed to be due to the basic architecture of the word recognition system (rather than being due to specific strategies). Pexman and Lupker (1999) argued that, if this account is correct, the two effects should occur simultaneously (i.e., in the same trial block) and both effects should be larger whenever there is increased opportunity for feedback to affect processing (i.e., when pseudohomophone foils are used). As predicted, Pexman and Lupker found that polysemy and homophone effects co-occurred and both were significantly larger with pseudohomophone foils than with pseudoword foils, supporting the feedback activation account.

One additional result that is consistent with Hino and Lupker's (1996) account comes from Pecher's (in press) examination of a different semantic factor: number of synonyms. Whereas polysemous words involve a many-to-one feedback mapping from the semantic units to the orthographic units (which helps increase the activation of the appropriate orthographic units), words with synonyms involve a one-to-many feedback mapping from the semantic units to the orthographic units. Thus, the feedback activation for a word with synonyms would tend to be dispersed to different orthographic representations, which should produce competition at the orthographic level. As a result, in contrast to the processing advantage created by polysemy, words with synonyms should be at a processing disadvantage. Pecher reported that responses were slower for words with synonyms (e.g., JAIL) than for words without synonyms (e.g., MILK) in both LDT and naming, and explained these results in terms of feedback processes.

The purpose of present paper was to provide a new examination of the feedback activation account. Polysemous words, like BANK, have a number of different, relatively distinct, meanings. Thus, according to the feedback activation account, these words create considerable semantic activation and, hence, more feedback activation for the orthographic and phonological units, producing faster responding. A similar situation should arise with any words that create relatively more semantic activation, regardless of whether that activation corresponds to several distinct meanings. In order to examine this prediction, we investigated the effect of number of features in LDTs and naming tasks.

Semantic features are attributes or characteristics that describe the meaning of a word. For instance, for the word LAMP, its semantic features might include such things as "is bright", "has light bulbs", "produces heat", "has a shade", etc. The notion that word meanings can be represented by semantic features has been controversial (e.g., Keil, 1989; Medin, 1989; Rips, 1989). That is, concept representations seem to involve much more than feature information; including such things as general world knowledge about relations between features, and heuristics like essentialism (the notion that things like lamps have "essences"). McRae, de Sa, and Seidenberg (1997; see also McRae, Cree, Westmacott, & de Sa, 1999) suggested, however, that featural representations do play an important role in at least the initial computation of word meaning. Based on the feedback activation account, it would be predicted that words with many features would produce

more semantic activation and, hence, more feedback to the orthographic and phonological units than words with few features. Thus, in LDTs and naming tasks, faster responding should be observed for words with a large number of features than for words with a small number of features.

The suggestion that word recognition may be faster for words with more semantic activation, or "richer" semantic representations, is not a new one. In previous research, effects of concreteness and/or imageability have been examined (e.g., Cortese, Simpson, & Woolsey, 1997; de Groot, 1989; James, 1975; Strain & Herdman, 1999; Strain, Patterson, & Seidenberg, 1995; Zevin & Balota, 2000), with results tending to show faster responding in LDTs and naming tasks for concrete or imageable words than for abstract words. It has been argued, in fact, that highly imageable or concrete words have richer semantic representations because they activate more semantic features than abstract words (Jones, 1985; Plaut & Shallice, 1993). According to the feedback activation account, however, activation of a larger number of semantic features should facilitate word recognition even when all of the stimuli are highly imageable. That is, even if all of the target words are concrete nouns, if some words activate more semantic features than others do, they should produce more rapid responding in word recognition tasks. Thus, there should be number of features (NOF) effects when concreteness and imageability have been controlled.

In this research we tested these predictions. Experiments 1A and 1B were LDTs, and 1C was a naming task.

Method

Participants

The participants in these experiments were undergraduate students at the University of Calgary. There were 40 participants in Experiment 1A, 38 in Experiment 1B, and 35 in Experiment 1C.

Stimuli

Words The word stimuli for Experiments 1A, 1B, and 1C, were selected from norms provided by Ken McRae (see McRae & Cree, in press). The McRae norms were collected by asking participants to list features for a large number of concrete nouns. Two sets of words were created: one set consisted of 25 words with low NOF and the other set consisted of 25 words with high NOF. These sets were matched on several dimensions. The mean values on these dimensions, for the selected sets of words, are listed in Table 1.

Foils There were 60 pseudowords presented in Experiment 1A and 60 pseudohomophones presented in Experiment 1B.

Procedure

On each trial, a letter string was presented in the centre of a 17-inch Sony Trinitron monitor controlled by a MacIntosh G3 and presented using PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993). In Experiments 1A and 1B, lexical-decision responses were made by pressing either the left button (labeled NONWORD) or the right button (labeled WORD) on a PsyScope response box. In Experiment 1C, naming responses were made into a microphone attached to a PsyScope response box.

Table 1: Mean Characteristics for Word Stimuli

Word characteristic	Low NOF words	High NOF words	Difference test $t(48)$
Number of features	12.00	20.40	-18.05**
Kucera & Francis (1967) frequency	10.80	14.32	<1
Subjective familiarity	3.84	3.97	<1
Number of meanings	1.08	1.07	<1
Word length	6.28	5.52	1.65
Number of syllables	1.80	1.56	1.25
Orthographic neighborhood size	3.00	3.64	<1

** $p < .001$

Experiment 1A – Results and Discussion

For this experiment, mean response latencies and mean error percentages are presented in Table 2. In all experiments, data were analyzed with subjects ($F1$ or $t1$) and, separately, items ($F2$ or $t2$) treated as random factors.

For high NOF words, response latencies were faster and there were fewer response errors (compared to responses for low NOF words) and, thus, there were significant NOF effects in both the latency analysis ($t1(39) = 2.95$, $p < .005$, $SE = 5.13$; $t2(48) = 1.40$, $p = .16$, $SE = 15.30$), and in the error analysis ($t1(39) = 2.66$, $p < .01$, $SE = 0.79$; $t2(48) = 1.17$, $p = .25$, $SE = 1.88$).

The results of Experiment 1A demonstrated that participants could more easily make word/nonword decisions for high NOF words than for low NOF words. According to the feedback activation account, this advantage was due to the additional semantic activation created by high NOF words. This additional semantic activation provided stronger feedback to the orthographic representation for the word presented, enhancing the activation of its orthographic units and speeding responding. In order to examine this NOF effect further, we used pseudohomophones as foils in Experiment 1B. According to the feedback activation account, these foils make lexical decisions more difficult because they require participants to set a higher criterion for orthographic activation. This leads to longer response times for both words and foils and increases the opportunity for feedback activation to affect responding. Thus, if the NOF effect is due to feedback activation from semantics to orthography, the effect should be larger in Experiment 1B.

Table 2: Mean Lexical Decision Latencies and Mean Error Percentages for Experiments 1A and 1B

	Experiment 1A (pseudoword foils)				Experiment 1B (pseudohomophone foils)			
Stimulus type	RT	Error %	RT effect	Error effect	RT	Error %	RT effect	Error effect
High NOF word	525	2.9			555	3.0		
Low NOF word	541	5.0	-16**	-2.1**	590	5.3	-35**	-2.3**
Foil	602	4.0			650	6.2		

**$\underline{p} < .01$

Experiment 1B – Results and Discussion

For this experiment, mean response latencies and mean error percentages are presented in Table 2.

As in Experiment 1A, response latencies were faster and there were fewer response errors for high NOF words, and so there was a significant NOF effect in the latency analyses ($\underline{t}1(37) = 5.01$, $\underline{p} < .001$, $\underline{SE} = 7.28$; $\underline{t}2(48) = 2.01$, $\underline{p} = .05$, $\underline{SE} = 21.30$), and in the error analysis ($\underline{t}1(37) = 2.98$, $\underline{p} < .005$, $\underline{SE} = 0.78$; $\underline{t}2(48) = 1.18$, $\underline{p} = .24$, $\underline{SE} = 1.95$).

In Experiment 1C we tested an additional prediction of the feedback activation account: because semantic activation also facilitates the activation of phonological units, high NOF words should also produce faster naming latencies.

Experiment 1C – Results and Discussion

For this experiment, mean naming latencies and mean error percentages are presented in Table 3.

Table 3: Mean Naming Latencies and Mean Error Percentages for Experiment 1C

Stimulus type	RT	Error %	RT effect	Error effect
High NOF word	525	0.3		
Low NOF word	555	1.4	-30**	-1.1*

*$\underline{p} < .05$, **$\underline{p} < .01$

For high NOF words, naming latencies were faster and there were fewer response errors, so there was a significant NOF effect in the latency analyses ($\underline{t}1(34) = 10.36$, $\underline{p} < .001$, $\underline{SE} = 2.96$; $\underline{t}2(48) = 2.09$, $\underline{p} < .05$, $\underline{SE} = 16.38$), and in the error analysis ($\underline{t}1(34) = 2.33$, $\underline{p} < .05$, $\underline{SE} = 0.45$; $\underline{t}2(48) = 1.41$, $\underline{p} = .16$, $\underline{SE} = 0.83$).

Again, responses were faster for words with high NOF. This suggests that semantic activation also provides strong feedback to the phonological units, facilitating naming responses.

Our 2 sets of words were not perfectly matched; there were slight differences between sets on several dimensions. To ensure that these differences were not the source of the observed effects, we conducted regression analyses. These analyses showed significant, unique effects of NOF for response latencies and response errors in Experiment 1A, response latencies (but not errors) in Experiment 1B, and naming latencies (but not naming errors) in Experiment 1C.

General Discussion

The present results demonstrate the influence of a previously unexamined semantic variable on visual word recognition. In the past, effects have been reported for concreteness and imageability (e.g., Cortese, Simpson, & Woolsey, 1997; de Groot, 1989; James, 1975; Strain & Herdman, 1999; Strain, Patterson, & Seidenberg, 1995; Zevin & Balota, 2000), and for polysemy (e.g., Borowsky & Masson, 1996; Gottlob et al., 1999; Hino & Lupker, 1996; Hino et al., 1998; Jastrzembski, 1981; Jastrzembski & Stanners, 1975; Kellas et al., 1988; Lichacz et al., 1999; Millis & Button, 1989; Pexman & Lupker, 1999; Rubenstein et al., 1970). The number of features effects reported here are independent of these effects. Our word stimuli were all concrete nouns, and were all nonpolysemous, differing only in terms of how many features participants ascribed to those words. Thus, our results provide support for the claim that it is the "richness" of a semantic representation that facilitates word recognition regardless of how that richness is created.

We have argued here that the NOF effects observed in our LDT and naming experiments (as well as a number of other semantic effects) support Hino and Lupker's (1996) feedback activation account. A key issue to address is to what extent other models of semantic effects, in particular, polysemy effects, could explain our NOF effects.

Alternative Explanations

Kawamoto et al. (1994) reported a successful simulation of polysemy effects in LDT using a model in which it was assumed that: (a) lexical decision performance is mainly based on activation of the orthographic units and (b) as a result of learning with their particular error-correction algorithm, weights for connections between orthographic units were enacted differently for polysemous and nonpolysemous words. Polysemy was captured in the model by having two different semantic patterns linked to a

single orthographic pattern. This inconsistent orthographic-to-semantic mapping created weaker connections between orthography and semantics. As a result, connections among orthographic units became more important in producing the appropriate orthographic activation for polysemous targets. In contrast, for nonpolysemous targets, semantic activation played a major role in producing the appropriate level of orthographic activation.

With respect to NOF effects, however, there would seem to be no reason why the number of features would affect the strength of either orthographic-to-semantic mappings or the connections among orthographic units. Neither our low nor high NOF words involved any orthographic-to-semantic inconsistencies. Thus, the model would have no obvious way to explain a NOF effect.

Borowsky and Masson (1996) successfully simulated their polysemy effects with a model in which it was assumed that lexical decisions are made on the basis of the "familiarity for a letter string's orthography and meaning" (p. 76). The model was a Hopfield network, and familiarity was assumed to be represented by the summed energy within the orthographic and meaning modules, with this energy reflecting the extent to which the network had settled into a basin of attraction. Energy was higher for polysemous words than for nonpolysemous words, due to proximity. That is, in the model, all the meaning-level units were initially set to +1 or –1 in a random fashion. Each unit was then updated until the network moved into a correct pattern. The distance (or the number of units to be changed) from the initial pattern to the correct pattern was probabilistically smaller when there were two correct patterns of activation (i.e., for polysemous words) than when there was only one correct pattern (i.e., for nonpolysemous words). Thus, the network moved into a basin of attraction more quickly for polysemous words than for nonpolysemous words, explaining the polysemy effect observed in LDT.

With respect to NOF effects, regardless of how many features a word has, it has only a single correct pattern of semantic activation. Thus, words with many features would not benefit from proximity like polysemous words do. Therefore, as with Kawamoto et al.'s (1994) model, this model would have no obvious way of explaining NOF effects.

It is possible that either of these models could be modified in a way that would allow them to explain NOF effects in LDT. In neither case, however, would the models provide as parsimonious an account as that provided by the feedback activation account. Further, in both cases, new assumptions would be needed to explain NOF effects in naming.

The results of the present experiments provide evidence that LDT and naming performance is faster for words with rich semantic representations, where richness is defined in terms of the number of semantic features activated. These effects suggest that word recognition performance will be best explained by fully-interactive models involving both feedforward and feedback activation.

Acknowledgments

This research was supported in part by grants to Penny M. Pexman and Stephen J. Lupker from the Natural Sciences and Engineering Research Council of Canada (NSERC). The authors thank Ken McRae for providing us with the feature norms, and Lorraine Reggin and Jodi Edwards for programming the experiments and testing participants.

References

Balota, D. A., Ferraro, R. F., & Connor, L. T. (1991). On the early influence of meaning in word recognition: A review of the literature. In P. J. Schwanenflugel (Ed.), *The psychology of word meanings* Hillsdale, NJ: Erlbaum.

Besner, D., & Joordens, S. (1995). Wrestling with ambiguity---Further reflections: Reply to Masson and Borowsky (1995) and Rueckl (1995). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 515-519.

Borowsky, R., & Masson, M. E. J. (1996). Semantic ambiguity effects in word identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 63-85.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, and Computers, 25,* 257-271.

Cortese, M. J., Simpson, G. B., & Woolsey, S. (1997). Effects of association and imageability on phonological mapping. *Psychonomic Bulletin and Review, 4,* 226-231.

de Groot, A. M. (1989). Representational aspects of word imageability and word frequency as assessed through word association. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 824-845.

Gottlob, L. R., Goldinger, S. D., Stone, G. O., & Van Orden, G. C. (1999). Reading homographs: Orthographic, phonologic, and semantic dynamics. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 561-574.

Hino, Y., & Lupker, S. J. (1996). Effects of polysemy in lexical decision and naming: An alternative to lexical access accounts. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 1331-1356.

Hino, Y., Lupker, S.J., Sears, C.R., & Ogawa, T. (1998). The effects of polysemy for Japanese katakana words. *Reading and Writing: An Interdisciplinary Journal, 10,* 395-424.

James, C. T. (1975). The role of semantic information in lexical decisions. *Journal of Experimental Psychology: Human Perception and Performance, 1,* 130-136.

Jastrzembski, J. E. (1981). Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon. *Cognitive Psychology, 13,* 278-305.

Jastrzembski, J. E., & Stanners, R. F. (1975). Multiple word meanings and lexical search speed. *Journal of Verbal Learning and Verbal Behavior, 14,* 534-537.

Jones, G. V. (1985). Deep dyslexia, imageability, and ease of predication. *Brain and Language, 24,* 1-19.

Joordens, S., & Besner, D. (1994). When banking on meaning is not (yet) money in the bank: Explorations in connectionist modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1051-1062.

Kawamoto, A. H., Farrar, W. T., & Kello, C. T. (1994). When two meanings are better than one: Modeling the ambiguity advantage using a recurrent distributed network. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 1233-1247.

Keil, F. C. (1989). *Concepts, kinds and cognitive development*. Cambridge, MA: MIT Press.

Kellas, G., Ferraro, F. R., & Simpson, G. B. (1988). Lexical ambiguity and the timecourse of attentional allocation in word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 601-609.

Kucera, H., & Francis, W. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Lichacz, F. M., Herdman, C. M., LeFevre, J., & Baird, B. (1999). Polysemy effects in naming. *Canadian Journal of Experimental Psychology, 53,* 189-193.

McRae, K., & Cree, G. S. (in press). Factors underlying category specific semantic deficits. In E. M. E. Forde & G. W. Humphreys (Eds.), *Category-Specificity in Brain and Mind.* East Sussex, UK: Psychology Press.

McRae, K., Cree, G. S., Westmacott, R., & de Sa, V. R. (1999). Further evidence for feature correlations in semantic memory. *Canadian Journal of Experimental Psychology, 53,* 360-373.

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General, 126,* 99-130.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist, 44,* 1469-1481.

Millis, M. L., & Button, S. B. (1989). The effect of polysemy on lexical decision time: Now you see it, now you don't. *Memory & Cognition, 17,* 141-147.

Pecher, D. (in press). Perception is a two-way junction: Feedback semantics in word recognition. *Psychonomic Bulletin and Review.*

Pexman, P. M., Lupker, S. J., & Jared, D. (2001). Homophone effects in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 139-156.

Pexman, P. M., & Lupker, S. J. (1999). Ambiguity and visual word recognition: Can feedback explain both homophone and polysemy effects? *Canadian Journal of Experimental Psychology, 53,* 323-334.

Piercey, C. D., & Joordens, S. (2000). Turning an advantage into a disadvantage: Ambiguity effects in lexical decision versus reading tasks. *Memory & Cognition, 28,* 657-666.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology, 10,* 377-500.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning.* Cambridge, England: Cambridge University Press.

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior, 9,* 487-494.

Rubenstein, H., Lewis, S. S., & Rubenstein, M. A. (1971). Evidence for phonemic recoding in visual word recognition. *Journal of Verbal Learning and Verbal Behavior, 10,* 645-657.

Rueckl, J. G. (1995). Ambiguity and connectionist networks: Still settling into a solution---Comment on Joordens and Besner (1994). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 501-508.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96,* 523-568.

Stone, G. O., Vanhoy, M., & Van Orden, G. C. (1997). Perception is a two-way street: Feedforward and feedback phonology in visual word recognition. *Journal of Memory and Language, 36,* 337-359.

Strain, E. & Herdman, C. M. (1999). Imageability effects in word naming: An individual differences analysis. *Canadian Journal of Experimental Psychology, 53,* 347-359.

Strain, E., Patterson, K., & Seidenberg, M. S. (1995). Semantic effects in single-word naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1140-1154.

Van Orden, G. C., & Goldinger, S. D. (1994). The interdependence of form and function in cognitive systems explains perception of printed words. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 1269-1291.

Zevin, J. D., & Balota, D. A. (2000). Priming and attentional control of lexical and sublexical pathways during naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 121-135.

Category learning without labels—A simplicity approach

Emmanuel Minos Pothos (e.pothos@ed.ac.uk)
Department of Psychology, University of Edinburgh; 7 George Square
Edinburgh, EH8 9JZ UK

Nick Chater (nick.chater@warwick.ac.uk)
Department of Psychology, University of Warwick;
Coventry, CV4 7AL UK

Abstract

In an extensive research tradition in categorization, researchers have looked at how participants will classify new objects into existing categories; or the factors affecting learning to associate category labels with a set of objects. In this work, we examine a complementary aspect of categorization, that of the spontaneous classification of items into categories. In such cases, there is no "correct" category structure that the participants must infer. We argue that the this second type of categorization, unsupervised categorization, can be seen as some form of perceptual organization. Thus, we take advantage of theoretical work in perceptual organization to use simplicity as a principle suitable for a model of unsupervised categorization. The model applied directly to similarity ratings about the objects to be categorized successfully predicted participants' spontaneous classifications. Moreover, we report evidence whereby perceived similarity is affected by spontaneous classification; this supplements the already substantial literature on such effects, but in categorization situation where the objects' classification is not pre-determined.

There are several situations in real life where novel objects can be spontaneously organized into groups. Consider a set of pebbles taken from a beach, or cloud patterns on a particular day, or just meaningless shapes shown onto a computer screen. This spontaneous classification can be appropriately labeled "unsupervised" because there are no "correct" categories the observer need to infer. By contrast, in supervised categorization, the learner (e.g., a child or someone learning a new language), has to infer what a category is by observing exemplars of the category and guessing their category membership (e.g., a child could be corrected for calling an apple an orange; through a process of corrective feedback, she would eventually learn to associate the appropriate objects with the category label "orange").

Supervised vs. unsupervised categorization

While there has been very little theoretical work on unsupervised categorization, this has not been the case for supervised categorization. Several models have been put forward, covering different intuitions about the cognitive mechanisms of supervised categorization. For example, in definitional accounts of concepts (e.g., Katz & Fodor, 1963), categories are characterized by necessary and sufficient conditions for an item to be a category member (see Pothos & Hahn, 2000, for a recent evaluation). In exemplar theories (e.g., Nosofsky, 1989), a concept is represented by a set of known instances of that concept; new instances are therefore assigned to different categories in terms of their similarity to the members of each category. In prototype theories assignment is also determined by a similarity process, but this time to the prototype of each category, where a category prototype encapsulates some measure of central tendency across the exemplars of the category (e.g., Homa, Sterling, & Trepel, 1981).

Despite the technical sophistication of this research, it does not cover the whole scope of categorization processes. Models such as the exemplar model or the prototype one could never be used to predict how a person would spontaneously classify a set of items. In fact, in an influential paper Murphy and Medin (1985) criticized models such as the above for failing to explain category coherence—why it is the case that certain groupings of items make better categories than others; for example, the categories of birds or cups are coherent, but a category consisting of dolphins born on Tuesdays together with pink tulips within 20 miles of London, and the Eiffel Tower would be nonsensical. Given that the exemplar or prototype models could not explain such observations, Murphy and Medin concluded that they are inadequate models of categorization (and thus made a case for the importance of general knowledge in categorization).

However, under the light of the present distinction between supervised and unsupervised categorization, it is not the case that the exemplar or the prototype modes are inadequate in that they fail to capture general knowledge effects. Rather, category coherence is a problem of unsupervised categorization, as it relates to how categories originate–a process which, necessarily, cannot be guided by a 'supervisor.'

To summarize this section, the distinction of categorization models into supervised and unsupervised serves the useful purpose of enabling a closer specifi-

cation of the type of results that we expect each model to be able to capture. Unsupervised models of categorization will fail in predicting how participants will classify a new instance into a set of existing categories; but such models could probably be used to ground a theory of category coherence. The converse applies to models of supervised categorization.

Previous work on unsupervised categorization

There has been an extensive experimental tradition on spontaneous classification, under the name of free sorting. However, the objective of free classification research is to identify the factors that appear to influence performance in sorting tasks, such as different types of instructions / experimental procedures and the structure of the stimuli (e.g., whether they are made of integral or separable dimensions, and the extent to which this affects the number of dimensions used in the classification task; e.g., Handel & Preusser, 1970; Wills & McLaren, 1998; Kaplan & Murphy, 1999). Thus, results from free sorting do not bear directly on the study of spontaneous classification, in the sense of actually predicting the classifications people are likely to come up with.

Trying to predict how objects are divided into groups has been a very frequently researched topic. While an exhaustive review of the different accounts by far exceeds the scope of the present work, we next discuss some of the qualifying factors of previous work with respect to its appropriateness for modeling unsupervised categorization.

Within machine learning and statistics, there is a long literature on clustering. There are two broad classes of clustering algorithms, agglomerative models and K-means ones. In the former case, for a set of N objects a hierarchy of clusters is produced whereby in the bottom level there are N clusters (a cluster for each object) and in the top level only one cluster (which includes all the objects). In the latter case, the number of clusters in which a set of objects is to be divided is set externally (this is why this approach is called "K-means"; for a review see Krzanowski & Marriott, 1995). In both approaches, knowledge of the number of groups sought is assumed; it must be pre-determined by the researcher. However, for a psychological model of unsupervised categorization we need to be able to predict both the number of categories and how the objects to be categorized are portioned into these categories within the same formalism.

This turns out to be an important limitation in terms of applying previous relevant modeling work in psychology directly to the problem of unsupervised categorization, as well. This applies, for example, to Ahn and Medin's Two Stage Model of Category Construction (Ahn & Medin, 1992), Michalski and Stepp's (1983) CLUSTER/2, and Anderson's rational categorization work (1991; additionally, Anderson's model is sensitive to order of presentation of the items to be categorized, so that his work is directed more towards dynamic aspects of categorization). This is not to criticize any of the excellent work cited above, but rather attempt to specify more precisely its modeling objective, with respect to how well it applies to unsupervised categorization.

Perhaps more directly relevant is Fisher's COBWEB (e.g., Fisher, 1996), which is based on the psychologically motivated principle of category utility (e.g., Corter & Gluck, 1992). Variants of the model can indeed determine the number of categories, as well as the way the items should be partitioned into the categories. However, three factors prevent its direct comparability to the present model. Firstly, category utility has been put forward to explain basic level categorization (e.g., Rosch & Mervis, 1985); the relation between basic level categorization is presently unknown. Secondly, COBWEB has been investigated—and to a large extent validated—as a statistical model, not a psychological algorithm. One of the differences between the two is that a psychological model is supposed to be founded on computational principles that make some statement about cognition. Finally, category utility assumes a representation of objects in terms of features; categorization predictions in this work are derived on the basis of empirically derived similarity information.

Perceptual organization and simplicity

Categorization and perceptual organization, albeit superficially dissimilar processes, are nevertheless quite interlinked. Clearly categorization depends on perceptual organization, as how we perceive a set of objects will by necessity determine how we will categorize them. However, there is also a very extensive research tradition on effects of categorization on perceptual organization, showing that the way we categorize a set of objects is likely to affect how we perceive them (e.g., Goldstone, 1994; Harnad, 1987; Schyns & Oliva, 1998). Thus, we could maybe usefully look for a principle in perceptual organization to ground our model of unsupervised categorization.

A very influential approach in perceptual organization is the simplicity principle (e.g., Pomerantz and Kubovy, 1986; Chater, 1999), according to which the perceptual system is viewed as finding the simplest perceptual organization consistent with the sensory input. In fact, the simplicity principle has been recently shown to be equivalent to the most influential alternative, the likelihood principle (Chater, 1996).

In a simplicity framework, the notions of "interpretation" and "encoding" are central. At an intuitive level, encoding of information results in some data; simplicity is just a strategy for choosing an interpretation for the data. If we have a sequence like "abababababab" we could interpret it as "5 x (ab)"; but, clearly, there are many alternative interpretations (e.g., "a, 2 x (baba), b"). According to simplicity, the preferred theory / in-

terpretation is the one that minimizes the sum of the (1) complexity of the theory and (2) the complexity of the data when encoded with the theory.

The simplicity model of unsupervised categorization

Full details of the model are given in Pothos & Chater (1998) and Pothos & Chater (in press). Here, we only attempt to qualitatively discuss the main features of the model.

There has been extensive research on the importance of information in categorization, other than similarity. However, there must be an important component of categorization research that is driven primarily by similarity as well. This would be particularly evident in the case of grouping novel objects, since there would be no a priori expectations for such objects. Also, incorporating general knowledge influences in models of categorization has been notoriously difficult. Thus, in this work we will restrict the simplicity model to a version whereby general knowledge effects are not taken into account.

We assume that the information encoded for a set of objects is information about how similar each object is to each other. A possible "interpretation" for this information is in terms of groups of categories; in other words, the cognitive system could attempt to recognize structure in the encoded similarity information that is best captured by dividing the objects into groups.

To determine which grouping is most suitable we need to consider the following terms:

code length for similarities in terms of grouping + code length for grouping (1)

code length for similarities without groups (2)

The simplicity principle will support the classification such that (1) is a lot less than (2).

Translating the above intuition into a computational model, we consider similarity information of the form (object A, object B) more or less similar to (object X, object, Y). The advantage of this approach is that the applicability of our categorization model is not restricted by representational assumptions for the objects to be categorized. For example, we can equally well apply the model, whether the items to be categorized are represented as bundles of features, points in some multidimensional space, or even simply in terms of pairwise similarities.

We define a group or a category as a collection of objects such that the similarities between any two objects in the group are greater than the similarities between any two objects between groups. In this way, the similarity relations that would have had to be specified without groups are reduced. For example, if we have objects A, B, and C, and we put objects A and B in one group, while object C is on its own, then by the above, this is equivalent to saying that the similarity between A and B is greater than between A and C, and B and C. In this way, we have an "operative" definition of a category.

Thus, with groups we have some information gain, or reduction in code length, since we do not need to specify as many similarity relations; this would be the "gain" associated with a classification. However, it will rarely be the case that all the specified similarity relations will be correct; in other words, a particular grouping might specify that objects A, B are more similar to objects X, Y, when in fact it is the other way around. Thus, the overall classification gain will be reduced by the costs of correcting the errors; there is an additional cost required to specify which is the particular classification used (for the actual formulae and derivations see Pothos & Chater, 1998).

Experimental investigation

We wish to illustrate the applicability of the model with empirically derived similarity information about the items to be categorized. This approach is consistent with a growing trend in categorization research to take into account the well documented similarity structure changes that take place as a result of categorization.

The simplicity model can be used to predict the classification that should be most psychologically intuitive to naïve observers for a set of objects. We can thus examine the extent to which the classifications spontaneously produced by naïve observers are compatible with the simplicity model predictions.

Materials

We used 11 items that varied along two dimensions (the physical space representation is shown is Figure 1; a 12th item had to be eliminated from analyses as it was not the same in the ratings and categorization tasks). The two dimensions defined the size of a square and the size of the filled-circles texture inside the square (see Figure 2 for an example). The stimuli were presented in a folder, printed individually on A4 paper in black ink for the categorization task, and on a 15'' Macintosh computer screen when participants were asked for similarity ratings.

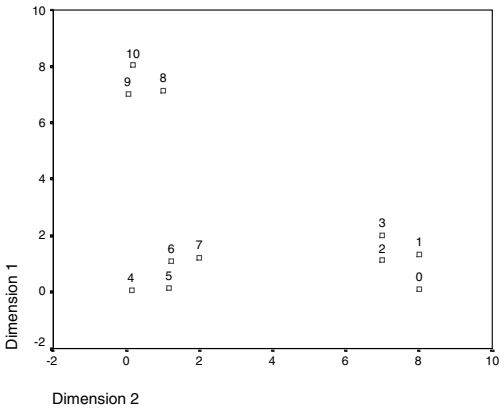


Figure 1: The parameter space representation of the stimuli.

Figure 2: An example of the stimuli used.

Procedure

29 University of Oxford students were paid for their participation. In the first part of the study, they received instructions saying that they were about to receive a set of items and that they would have to divide them into groups "in a way that seems intuitive and natural, so that more similar items end up in the same group." They were also told that although there was no limit on how many groups they could use, they should not use more that what they thought necessary. The order in which the stimuli were arranged in the folder was randomized for each participant.

After participants had classified the items, they performed the ratings task on a computer. They were instructed that they were about to see the items of the first part in pairs and that their task was to indicate the similarity between the items in each pair on a 1 to 9 scale, where a "1" would correspond to most similar items and a "9" to items that were most different. In particular, for each pair, the first item was presented for one and a half seconds, then there was a fixation point for 250ms, the second item appeared for one and a half seconds, a blank screen for 250ms, and a 1–9 ratings scale. The order in which each item appeared in a pair was counterbalanced so that we had two ratings per participant for each pair. Two randomized different orders were used for the ratings part of the experiment.

Results and Discussion

The similarity ratings were averaged into a large similarity matrix for all the items. This matrix was made symmetrical across the diagonal by using the arithmetic mean and also self-similarities were set to 0 (corresponding to maximum similarity). The simplicity model predictions were computed on the basis of these ratings. The best compression categorization involved three groups, with items 0–3, 4–7, and 8–10 in each group (item labels correspond to Figure 1).

In order to determine whether some of the observed categorizations were more likely than others we identified all the distinct categorizations produced by participants ("distinct" solutions), as well as the number of times participants divided the items in the way predicted by the simplicity principle. If there had been no preference for any particular categorization, we assumed that all distinct solutions would have been produced with a roughly equal frequency, given by the ratio (total number of groupings) / (number of distinct groupings). Using chi-square tests we can then examine whether the frequency of any of the classifications produced would be different from that computed by chance. This was the case only for the classification predicted by the model ($chi2(1) = 84.8$, $p<.001$; the frequency of this categorization was 11 times, out of 29).

To obtain some insights into participants' performance, we employed a non-metric MDS procedure to construct a putative internal spatial representation of the items; such a procedure is not related to the application of the simplicity model (which operates directly on the similarity ratings). Figure 3 shows the resulting MDS solution (all MDS procedures run with Euclidean metric). The three groups in the MDS solution correspond exactly to the three groups in Figure 1–but the items within each cluster are effectively indistinguishable in the internal space.

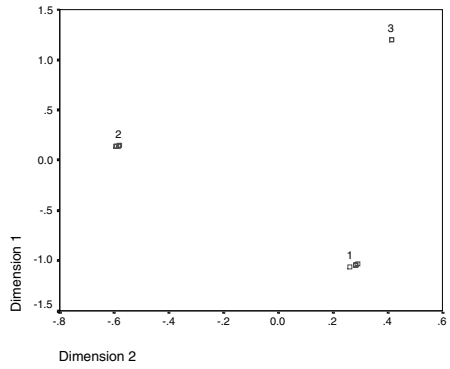


Figure 3: Labels items 0–3, "2" items 4–7, and "3" items 8–10, where item labels refer to Figure 1.

We next divided participants into homogeneous groups, and examined these groups individually. To do this, we looked at the groupings produced in the first part of the experiment, and then classified these groupings themselves (using the Rand index as a measure of the similarity between pairs of groupings, and the simplicity model as the clustering procedure). There were two main groups of categorizations, call them Group A (which contained the best compression solution; five different categorizations, that were produced by 14 out of the 29 participants) and Group B (nine solutions from 13 participants), as well as a smaller group which we shall not consider further (two other categorizations, from two participants).

We then separately considered the similarity ratings of participants whose groupings were in Groups A and B. The MDS procedure for Group A resulted in a spatial arrangement of the stimuli, identical to that shown in Figure 3. Figure 4, the MDS solution for Group B participants, is dramatically different; although some aspects of the nearest-neighbors structure seem to have been somewhat preserved (so that, for instance, points that were close to each other originally are still close to each other) the overall arrangement has been distorted so as to no longer reflect the obvious three groups category structure present in the Group A representation of the stimuli. In conclusion, it looks as if people who identified the best compression categorization (Group A), subsequently rated the similarity of different stimuli with each other in a way fully compatible with this category structure. This finding constitutes the first evidence that unsupervised classification affects the perceived similarity structure of a set of objects (see, e.g., Goldstone, 1995; Goldstone, Steyvers & Larimer, 1996 for corresponding evidence in supervised classification, that is categorization processes whereby categories are pre-specified). Future research will extend the present methodology to examine the extent to which simplicity might always be optimized with respect to how different individuals perceive the similarity structure of a set of objects.

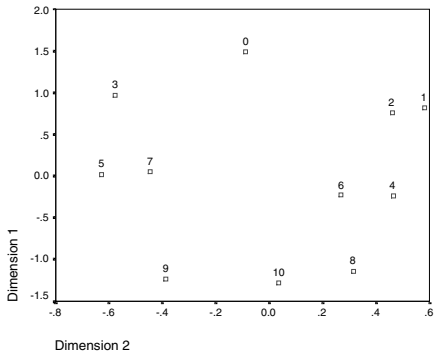


Figure 4: MDS solution for Group B participants.

To summarize, analysis of the similarity ratings of the stimuli confirmed the predictions of the simplicity model. Moreover, inspection of the MDS solutions showed that the categorization appears to have influenced similarity judgments, implying that perceived similarity may be affected by unsupervised classification.

References

Ahn, W. & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science*, 16, 81-121.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98, 409-429.

Chater, N. (1996). Reconciling Simplicity and Likelihood Principles in Perceptual Organization. *Psychological Review*, 103, 566-591.

Chater, N. (1999). The Search for Simplicity: A Fundamental Cognitive Principle? *Quarterly Journal of Experimental Psychology*, 52A, 273-302.

Corter, J. E. & Gluck, M. A. (1992). Explaining Basic Categories: Feature Predictability and Information. *Psychological Bulletin*, 2, 291-303.

Fisher, D. (1996). Iterative optimization and simplification of hierarchical clusterings. *Journal of Artificial Intelligence Research*, 4, 147-179.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, 123, 178-200.

Goldstone, R. L., Steyvers, M., & Larimer, K. (1996). Categorical perception of novel dimensions. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Handel, S. & Preusser, D. (1970). The free classification of hierarchically and categorically related stimuli. *Journal of Verbal Learning and Verbal Behavior*, 9, 222-231.

Harnad, S. (Ed.) (1987). *Categorical Perception*. Cambridge: Cambridge University Press.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory*, 7, 418-439.

Kaplan, A. & Murphy, G. L. (1999). The acquisition of category structure in unsupervised learning. *Memory & Cognition*, 27, 699-712.

Katz, J. & Fodor, J. A. (1963). The Structure of a Semantic Theory. *Language*, 39, 170-210.

Krzanowski, W. J. & Marriott, F. H. C. (1995). *Multivariate Analysis, Part 2: Classification, Covariance Structures and Repeated Measurements*. Arnold: London.

Michalski, R. & Stepp, R. E. (1983). Automated construction of classifications: conceptual clustering versus numerical taxonomy. *IEEE Transactions on pattern analysis and machine intelligence*, Vol. PAMI-5, 396-410.

Murphy, G. L. & Medin, D. L. (1985). The Role of Theories in Conceptual Coherence. *Psychological Review*, 92, 289-316.

Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Journal of Experimental Psychology: Perception and Psychophysics*, 45, 279-290.

Pomerantz, J. R. & Kubovy, M. (1986). Theoretical Approaches to Perceptual Organization: Simplicity and Likelihood principles. In: K. R. Boff, L. Kaufman & J. P. Thomas (Eds.), *Handbook of Perception and Human Performance, Volume II: Cognitive Processes and Performance*, 1-45. New York: Wiley.

Pothos, E. M. & Chater, N. (1998). Rational Categories. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, 848-853, LEA: Mahwah, NJ.

Pothos, E. M. & Hahn, U. (2000). So concepts aren't definitions, but do they have necessary *or* sufficient features?. *British Journal of Psychology*, 91, 439-450.

Pothos, E. M. & Chater, N. (in press). Basic Categories by Simplicity. In M. Ramscar, U. Hahn, E. Cambouropoulos, & H. Pain (Eds.) *Similarity and Categorization*. Oxford: Oxford University Press.

Rosch, E. & Mervis, B. C. (1975). Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology*, 7, 573-605.

Schyns, P. G. & Oliva, A. (1999). Dr. Angry and Mr. Smile: When categorization flexibly modifies the perception of faces in rapid visual presentations. *Cognition*, 69, 243-265

Wills, A. J. & McLaren, I. P. L. (1998). Perceptual learning and free classification. *Quarterly Journal of Experimental Psychology*, 51B, 235-270.

Neural Synchrony Through Controlled Tracking

Dennis Pozega (dpozega@engmail.uwaterloo.ca)
Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1 Canada

Paul Thagard (pthagard@uwaterloo.ca)
Department of Philosophy, University of Waterloo, Waterloo, ON N2L 3G1 Canada

Abstract

We present a model for generating a kind of neural synchrony in which the individual spike trains of one neuron or group of neurons closely match the spike trains of another. This kind of neural synchrony has been observed in animals performing auditory, visual and attentional information processing tasks. Our model is realized in a system of functionally identical, refractory spiking neurons. Larger systems with more sophisticated information processing capabilities can be constructed from aggregated instances of the basic network.

Introduction

Recently researchers ranging from neurobiologists (e.g. Ritz & Sejnowski, 1997) to computer scientists (e.g. Shastri, 1999) to psychologists (e.g. Hummel & Holyoak, 1997; Sougné, 2000) have studied the thesis that the synchrony of neural activity is one means by which bits of information are aggregated into the larger wholes necessary for complex information processing. In brief, neural synchrony is thought to implement the 'dynamic binding' of certain representations in the brain. It has also been suggested that neural synchrony is part of the process by which consciousness emerges from distributed brain activity.

But which neural dynamics are actually being synchronized? The answers provided by a decade of intensive research have generally all been variations on three basic themes. The first approach to synchrony intends the term to refer to coherence between 'oscillations' observed in the brain. In these cases it is the aggregated neural activity generated by the whole of a particular neural population that is deemed to be oscillating, not individual cortical neurons. Ritz and Sejnowski (1997) provide an excellent review of the conditions under which this mode of synchrony occurs, as well as its reputed significance in information processing tasks. Two representative examples include the observation of oscillatory synchrony in the brains of cats completing a sensorimotor task (Roelfsema, Engel, König & Singer, 1997), and its generation in an artificial neural network modeling the brain's solution of a figure-ground segregation problem (Sporns, Tononi & Edelman, 1991). Engel et al. have even speculated that oscillatory synchrony may be instrumental in bringing representations to consciousness, based on results showing that this mode of synchrony correlates with perceptual awareness in cats (1999).

Synchrony is understood elsewhere as correlated quasiperiodic activity occurring at a constant phase offset with respect to a persistent background oscillation. Different groups of neurons periodically become active during different phases of the background oscillation; neurons that become coactive during the same subperiods are said to exhibit phase synchrony. Phase synchrony allows the expression of phase-coded representations: representations in which information is coded in the relative timing of quasiperiodic neural activity. Gerstner, Kempter, van Hemmen and Wagner (1999) use empirical evidence for phase-coding's involvement in the sound source localization task in barn owls to build a successful mathematical simulation of the process. As well, Jensen and Lisman argue that data from psychological experiments and rat EEGs support accounts of short term memory capacity (1998) and position reconstruction in rats (2000), respectively, framed in phase-coding terms.

Phase synchrony is also exploited in a number of influential models of cognitive processes. Shastri, in his SHRUTI model of inference and reasoning, uses phase synchrony to bind neurons corresponding to role-filler words; e.g., John, with other neurons corresponding to specific roles in propositions; e.g., X in "X sees Y" (Shastri & Ajjanagadde, 1993; Shastri, 1999). Similarly, phase-coded bindings of roles to role-fillers underlie Hummel and Holyoak's IMM and LISA models of analogy formation (Hummel, Burns & Holyoak, 1994; Hummel & Holyoak, 1997).

Finally, there is a third synchrony phenomenon that we have termed 'spike train synchrony.' This synchrony is present when strong correlations exist between the individual firing times of different neurons or groups of neurons. There is no need that these firing times be quasiperiodic, as in phase synchrony. Nor is it generally possible to explain the aggregate synchronized activity as oscillatory; in fact, the mean overall activity of the neurons involved can remain close to constant except over very short time intervals.

DeCharms and Merzenich (1996) observed the spike train synchrony of neurons in the brains of anesthetized marmoset monkeys responding to a pure tone stimulus. For the duration of each tone, the firing patterns of selected neural regions became correlated, even though their mean firing rates remained unchanged. In addition, this correlation disappeared when the tone ended and was absent before it began. Tightly correlated spike trains have also been reported for neighbouring cells in two early vision regions: the retinal ganglia (Meister, 1996) and the lateral geniculate nucleus (Alonso, Usrey & Reid, 1996). More recently, Steinmetz et al. (2000) have found that certain somatosensory neurons in monkeys increase the correlation of their spike trains when performing visual and tactile tasks requiring increased attention.

All three modes of synchrony outlined here have been empirically observed under conditions which suggest a role for them in specific information processing tasks. Furthermore, previous research, as cited above, has resulted in working models demonstrating how the first two types, oscillatory and phase synchrony, may be generated. The design of computational simulations targeting the generation of synchrony at the level of individual spike trains, however, has received little attention.

Simulation-based research of this type should help to elucidate the structural and functional aspects of the brain necessary for spike train synchrony as it has been observed. Moreover, even in the absence of an artificial system mirroring the exact architecture of the brain, spike train synchrony models should provide insights and help researchers test hypotheses concerning information processing tasks whose realizations in the brain presumably require this form of synchrony. Advances generated by analogous models depending on other synchronies—Shastri's model of logical reasoning, for example—corroborate this claim.

In this paper we present an artificial neural network designed to exhibit spike train synchrony. The neurons in the network are all functionally identical, refractory neurons, and the connections between them are all of the same, standard type.

The controlled tracking network displays a simple behaviour: the *clone neuron* copies or 'tracks' the spike train of the *primary neuron*. The copying process is selective, meaning that it stops and starts in response to signaling from two *actuator neurons*. In this way, the *clone neuron* can be made to fall in and out of synchrony with the *primary neuron*.

The design of the network was completed in two stages. In the first, a basic network was built in which the clone neuron was made to copy the activity of the primary neuron at all times. Pausing or halting the copying is not allowed in this network. These operations were implemented in the second stage of network design, in which the control component of the system was integrated. Finally, in the last section, we highlight the contributions of the model to understanding the generation of biological spike train synchrony and its role in information processing. This includes a discussion of the merits of our model's representational capabilities over those used in other influential modeling approaches.

Mathematical Fundamentals

This section describes the dynamics of the neurons we use to build our spike-tracking networks. In brief, our neurons function like the Spike Response Model neurons developed in Gerstner (1999), with only slight modifications. We review the defining equations of our neurons here, drawing heavily from Gerstner's formulations.

The total membrane potential u_i of each spiking neuron i is given by:

$$u_i(t) = \sum_{t_i \in F_i} \eta(t - t_i) + \sum_{j \in \Gamma_i} \sum_{t_j \in F_j} w_{i,j} \varepsilon_{i,j}(t - t_j) . \quad (1)$$

The last term of this equation quantifies the contributions to the membrane potential of neuron i due to excitations and inhibitions from the set Γ_i of neurons with efferent connections to i. Each such neuron j will contribute to i's membrane potential due to post-synaptic potentials ε seen coming from j across the synapse connecting to i. The function $\varepsilon_{i,j}(s)$ equals the excitatory post-synaptic potential seen at post-synaptic neuron i at a time of s seconds after the firing of a pre-synaptic neuron j. The set F_i represents the set of all individual spike times t_i of the neuron i; likewise, F_j and t_j for neuron j. Finally, the constant $w_{i,j}$ represents the strength of the connection from neuron i to neuron j. In summary, the last term of Eq. (1) sums the contributions to membrane potential due to incoming post-synaptic potentials, after scaling these contributions by the appropriate connection weights.

The function $\varepsilon_{i,j}(s)$, introduced above is defined by:

$$\varepsilon_{i,j}(s) = \left[\exp\left(-\frac{s - \Delta_{i,j}}{\tau_m} \right) - \exp\left(-\frac{s - \Delta_{i,j}}{\tau_s} \right) \right] H(s - \Delta_{i,j}) , \quad (2)$$

where τ_m and τ_s are time constants determining the shape of the post-synaptic pulse, $\Delta_{i,j}$ is the propagation time of the electric potential signal between the beginning and end of a connection from i to j (also called the axonal delay or 'length') , and H is the Heaviside unit step function:

$$H(t) = \begin{cases} 0 & \text{for } t < 0 \\ 1 & \text{otherwise} \end{cases} \quad (3)$$

The left term on the right hand side of equation (1) accounts for the response of neuron i to its own previous spikes. This term quantifies the *refractoriness* of neurons; i.e., the decreased capacity of a neuron to spike soon after it has just spiked. Refractoriness is modeled as a short-term, decaying, inhibitory signal:

$$\eta(s) = -\eta_0 \exp\left(-\frac{s}{\tau_r}\right) H(s) - K[H(s)]H(\delta^{abs} - s) \quad (4)$$

The variable s is the time since a previous spike, while the parameters η_o and τ_r scale the amplitude and the decay rate, respectively. *H(s)* remains the Heaviside unit step function as discussed above. Finally, the constant K represents an arbitrarily large number and the parameter δ^{abs} designates the duration of absolute refractoriness. This represents the length of time after a spike during which a second spike is physically impossible.

A neuron spikes each time that its membrane potential u_i exceeds a threshold value ϑ, provided that u_i is increasing from a subthreshold value on the last discrete time step. Each spike results in a new spike time t_i being added to the set F_i of all spike times for neuron i.

Tracking Network

In this section, we describe a network where one neuron's spike train, the clone neuron, closely duplicates the spike train of another neuron, the primary neuron, at all times. We call this network the *tracking network*. Fig. 1 illustrates the architecture of our tracking network, in which auxiliary neurons serve to enable a clone neuron to synchronize its firing with the primary neuron. The complication of the network, primarily due to the many feedback connections, is necessary in a tracking system composed of refractory neurons. Refractoriness causes a transient desensitization of a neuron to incoming inputs shortly after that neuron's firing. Consequently, a pulse in an input signal will not elicit the same response from a refractory neuron that has just fired as it will from one that has not yet fired. This means that simply connecting the primary neuron to the clone neuron will not produce reliable tracking because the clone neuron will be biased to respond differently to different post-synaptic pulses sent from the primary neuron. The clone neuron will be slower to spike in response to an incoming superthreshold pulse coming a short time after the last one, in comparison to one coming a long time after the last one.

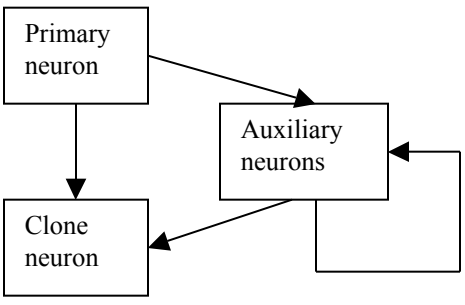


Figure 1. Spike tracking system.

To restore the efficacy of the external input to a refractory neuron, which in our system corresponds to the primary neuron's outgoing post-synaptic pulses, we construct an auxiliary signal that provides a second input to the clone neuron. This signal is implemented using the output pulses of *auxiliary neurons* connected efferently to the clone neuron. These auxiliary neurons are coerced to fire in such a way that the potential pulses they output individually to the clone neuron add up to a negative approximation of the η refractory term of Eq. (1).

Mathematically, each connection from an auxiliary neuron to the clone neuron results in an additional term in the expansion of the double summation of Eq. (1). Therefore, for the set A of auxiliary neurons to successfully mitigate refraction, the mathematical constraint to be satisfied is specified by Eq. (5) in the appendix. The variables $w_{c,a}$ and $\Delta_{c,a}$ represent the auxiliary to clone neuron connection strengths and time delays, respectively, while t_a represents the firing times of the auxiliary neurons. These three sets of variables remain to be specified, in addition to the total number of auxiliary neurons to be used. To determine acceptable values for these unknowns we first perform the following simplifying analysis. The analysis solves the approximation problem for only a limiting case, but we go on to show that straightforward modifications lead to a more general solution.

For now we assume a single, isolated refractory incident corresponding to a single, isolated firing of the clone neuron at time $t_c = 0$. Second, we require that pulses from the auxiliary neurons arrive instantaneously at the membrane of the clone neuron; i.e., we set the connection time delays $\Delta_{c,a}$ to 0. Finally, we tentatively prohibit each auxiliary neuron from firing more than once. These conditions allowed us to reformulate the constraint of Eq. (5) using Eq. (6) in the appendix. In graphical terms, this simplified constraint commits us to build up an approximation to the refractory function η by summing time-shifted and scaled postsynaptic pulses ε. The pulses are vertically scaled according to the connection strengths $w_{c,a}$. Their time shifting is specified by the times of firing t_a of the auxiliary neurons.

For convenience we chose to limit the maximum relative error of our particular approximation to less than 5%. Better approximations are possible, but as error decreases, the number of required auxiliary neurons increases, making the network larger. For our error tolerance of 5% we were able to construct a feasible solution set of values t_a and $w_{c,a}$ using 23 auxiliary neurons.

The set of solution values t_a indicate the time each auxiliary neuron must spike, *relative to the clone neuron*, in order to mitigate refractoriness. To force the neurons to spike at these times, we apply the same external input signal that the clone neuron experiences

to each of the auxiliary neurons, with a time delay of length t_a, the firing offset time for that neuron. In physical terms this means connecting the primary neuron to each auxiliary neuron. The strengths $w_{a,p}$ of these primary to auxiliary connections should be identical to the strength $w_{c,p}$ of the single primary to clone neuron connection. The lengths $\Delta_{a,p}$ of the primary to auxiliary connections, however, must equal the primary to clone neuron connection length *plus* the firing time offset t_a calculated for the particular auxiliary neuron *a* in question.

Returning to the assumptions we made earlier, the system thus connected is only strictly guaranteed to mitigate the clone neuron's refractoriness due to its very first spike. This is because the clone neuron was assumed to fire only once in isolation. To compensate for the refractoriness following *all* spikes, the auxiliary neurons need to spike—and with a precise time lag of t_a—not just the first time that the clone neuron spikes, but *every* time it spikes.

Though all the auxiliary neurons share the same external input as the clone neuron, they will fail to spike as required (i.e., every time the clone neuron spikes) because the clone neuron now has an additional input signal compensating for its internal refractoriness. It is no longer desensitized due to the refractoriness following from its first spike, but the auxiliary neurons still are.

We remedy this problem by feeding an additional input into each of the auxiliary neurons. Each of these additional input signals should be identical to the refractoriness compensation signal for the clone neuron, except that, as before, the signals should arrive with a time lag or delay of t_a specific to the auxiliary neuron *a* in question.

In terms of implementation, this translates into an additional bundle of connections to each auxiliary neuron. Each bundle consists of a set of connections leading from *each* auxiliary neuron to *one* of those auxiliary neurons. Therefore, if there are N auxiliary neurons, N feedback bundles are required, each containing N unique connections. This makes for N^2 auxiliary neuron feedback connections in total.

We should emphasize that we have been ignoring the absolute refractory component of the refractoriness signal corresponding to the second term in Eq. (2). Physiologically the absolute refractoriness of neurons could never be compensated for in biological networks in any event, because it arises from a fundamental electrochemical constraint on the availability of certain molecules (Paul, 1975). Furthermore, because all our neurons are identical, a clone neuron would never be expected to track a spike which would land itself in the absolute refractory period following a previous spike: the primary neuron would not be capable of producing such a spike train.

We tested the capacity of the clone neuron to copy the spike train of the primary neuron, as connected in the tracking network described above. The network was simulated using the SpikeSim program we designed in Java. As input to the tracking system, spike trains were evoked from the primary neuron that corresponded to an exponential distribution of spikes with a mean of 30 time steps between spikes.

The results indicate that missing spikes and bursts of spikes at inappropriate times ('ringing') seem to be the only symptoms of inaccurate tracking. In biological systems such aberrations could be explained by molecular and thermal noise.

A very strong cross correlation over a set of ten trials was observed between the spike trains of the primary and clone neuron. We found that the spike train of the clone neuron slightly lags the spike train of the primary neuron. A non-zero time lag is inevitable because it takes time for the post-synaptic pulses ε to peak. Nevertheless, it should not be of too much concern because the signals can be time-shifted so they coincide as seen from a third neuron's perspective. Setting appropriate connection lengths for the primary to third neuron connection and the clone to third neuron connection will implement the necessary shifting.

Discussion

We have presented a new computational model for synchrony generation, a model that implements controlled spike train tracking of one neuron by another. In this closing discussion, we investigate a possible variation of the model promising greater physiological plausibility. We then move on to compare the merits of this model to others in the field; namely, Shastri's and Hummel's. Finally, we argue that our basic model and its variations will likely prove helpful in the effort to develop larger scale simulations of cognitive processes.

To build the controlled spike tracking system, we constructed connections between the auxiliary neurons and the clone neuron of length (or time delay) zero. The connections between the primary neuron and the clone neuron, however, are variable in length and all non zero, in order to implement time lag delays. As a result of this configuration all the auxiliary neurons and clone neurons end up generating identical but time-shifted spike trains: these neurons fire in a wave-like or 'follow-the-leader' type manner.

Theoretic considerations suggest that an operationally identical network can be implemented in a way that is more physiologically realistic. Instead of implementing the time delays through lengthening the connections from the primary to auxiliary neurons, we can set these constant and stagger the lengths of two sets of connections: those in the feedback bundles, and those between the auxiliary neurons and the clone neuron. This removes the need for problematic zero-length connections across which electrical pulses would presumably need to instantaneously travel, while preserving the critical time-delay dynamics of the

refractoriness compensation signal. In this revised setup, the auxiliary neurons and the clone neuron would all fire identical spike trains, just as before, but with no time lag asynchrony.

We also reason that if auxiliary neurons exist in the brain to mitigate for refraction, they likely play roles in other brain circuits as well. If this is so, their mutually synchronous firing would help to maintain signal or information synchrony in all of these circuits. Such synchrony would not evolve from the original controlled tracking network, but would characterize the auxiliary neurons in the modified variation, as described above. In short, there are several reasons for believing that the proposed network variation should be superior to the original controlled tracking network.

The artificial neural networks whose results were presented above were simulated with the assumption of no noise. In later trials we investigated the effect of introducing noisy dynamics to the Spike Response Neuron model defining the neurons in our network. More specifically, the firing thresholds were made noisy by adding to the membrane threshold function a Gaussian random variable with mean of zero and standard deviation proportional to the degree of desired noisiness. This method of introducing noise is one of the standard ones discussed in Gerstner (1999).

When simulated, noisy neurons resulted in severely poor tracking performance, even when the noisiness was kept low. The largest errors in tracking were bursts of extraneous spikes occurring after a single premature firing. This initial firing was in turn due to a transient lowering of the membrane threshold of a single neuron when that neuron's membrane potential was close to the average threshold.

We expect that poor tracking under noise conditions can be mostly eliminated by a redesign of the feedback connections between auxiliary neurons, one which would not modify their basic role in the network. Our ideas are still only in an early stage of development, however. For now we have to concede that our system's spike tracking performance is poor under noisy conditions.

Future research could investigate the implications of the unique information encoding and processing properties our system possesses by nature of its design. For instance, the system allows for more versatile synchronic neural coding than that available in networks used in Shastri's SHRUTI model of logical reasoning or Hummel's LISA or IMM models of analogy formation. In these systems individual neurons are active for a maximum of one portion or 'phase' of each background oscillatory cycle, during which they may fire either alone or more generally in synchrony with a group of other neurons that are also active only during that phase. During other periods of the background oscillatory cycle all these neurons are inactive.

The neurons in the controlled spike-tracking system presented here can be made to synchronize in this way, but they are capable of more sophisticated synchronic dynamics as well, as the simulation results demonstrated. Through the control interface, variable length periods of transient synchrony were elicited from our system on demand. Furthermore, as will be demonstrated below, the clone neuron can be made to regularly switch between synchrony with different groups of neurons. In contrast, the other three models require a neuron to synchronize with only one group of neurons and only then during short, periodic time windows of constant width.

Another significant difference lies in the firing patterns during periods of synchrony. In the SHRUTI, LISA and IMM-based networks, the firing of a neuron within its interval of synchrony is only described in terms of its overall firing rate or 'level of activity.' But in controlled spike tracking systems, significant *subcoding* can take place within time intervals of synchronized activity. By subcoding we mean that additional information can be stored in the relative timing of spikes within each period or instance of neural synchrony. For the other models noted, supporting subcoding within periods of synchrony would effectively require system redesign. Consequently, we claim that our tracking system allows for more detailed elaboration of representations during synchronous neural firing.

Finally, we suggest that sophisticated neural networks for simulating higher level cognitive processes could be designed using the controlled tracking system as a reusable component, a repeated building block. Such networks would broaden our understanding of the processes they model, just as SHRUTI and LISA have provided insights into the reasoning and analogy-forming processes they were built to simulate.

Imagine a network with six inputs. Three of these are generated by three primary neurons, each of which we assume signals a separate and unique stream of representations coded within its spike train. The other three inputs are provided by three *switch neurons*, each corresponding to one of the three primary neurons. The network takes advantage of these inputs, and four spike tracking subsystems embedded within it, to implement a relatively complex behaviour: on receiving a single spike from one of the three switch neurons, the network begins to track the spike train of the corresponding primary neuron. When a different switch neuron spikes, the network switches inputs and starts tracking a different primary neuron. In effect, this network implements the ability to reorganize arbitrary parts of neural signals, corresponding to different temporally coded representations, in sophisticated ways. Such reorganization or 'splicing' mechanisms could prove quite useful for a number of information processing tasks. Compressions of

sequences of elaborate representations into more succinct streams could be one such task.

This network demonstrates the potential increase in information processing sophistication that emerges from constructing larger networks within which simple spike tracking subsystems work in coordination. We hope that the application of this kind of design approach will contribute to the future development of neural network models that carry out information processing tasks similar in complexity to those that the brain performs. Moreover, by shedding light on the internal mechanics of the processes involved, such models might not only demonstrate or mimic *what* the brain does but also help to further clarify just *how* it does it.

Acknowledgments

We thank Brandon Wagar and Sid Fingerote for helpful discussions. This work was funded by NSERC.

References

Alonso, J.-M., Usrey, W. M., & Reid, R.C. (1996). Precisely correlated firing in cells of the lateral geniculate nucleus. *Nature*, 383, 815-819.

deCharms, R. C. & Merzenich, M. M. (1996). Primary cortical representation of sounds by the coordination of action-potential timing. *Nature*, 381, 610-613.

Engel, A.K., Fries, P., König, P., Brecht, M., & Singer, W. (1999). Temporal binding, binocular rivalry, and consciousness. *Consciousness and Cognition*, 8, 128-151.

Gerstner, W. (1999). Spiking neurons. In W. Maass & C. M. Bishop (Eds.), *Pulsed neural networks* (pp. 3-53). Cambridge, Mass.: MIT Press.

Gerstner, W., Kempter, R., van Hemmen, J. L., & Wagner, H. (1999). Hebbian learning of pulse timing in the barn owl auditory system. In W. Maass & C. M. Bishop (Eds.), *Pulsed neural networks* (pp. 353-377). Cambridge, Mass.: MIT Press.

Hummel, J. E., Burns, B., & Holyoak, K. J. (1994). Analogical mapping by dynamic binding: preliminary investigations. In K. J. Holyoak & J. A. Barnden (Eds.), *Advances in connectionist and neural computation theory: volume 2* (pp. 416-445). Norwood, New Jersey: Ablex.

Hummel, J. E. & Holyoak, K. J. (1997). Distributed representations of structure: a theory of analogical access and mapping. *Psychological Review*, 104(3), 427-466.

Jensen, O. & Lisman, J. E. (2000). Position reconstruction from an ensemble of hippocampal place cells: contribution of theta phase coding. *Journal of Neurophysiology*, 83, 2602-2609.

Jensen, O. & Lisman, J. E. (1998). An oscillatory short-term memory buffer model can account for data on the Sternberg task. *The Journal of Neuroscience*, 18(24), 10688-10699.

Meister, M. (1996). Multineuronal codes in retinal signaling. *Proceedings of the National Academy of Science USA*, 93, 609-614.

Paul, D. H. (1975). *The physiology of nerve cells* (pp. 57-8). Oxford: Blackwell Scientific.

Ritz, R. & Sejnowski, T. J. (1997). Synchronous oscillatory activity in sensory systems: new vistas on mechanisms. *Current Opinion in Neurobiology*, 7, 536-546.

Roelfsema, P. R., Engel, A. K., König, P., & Singer, W. (1997). Visuomotor integration is associated with zero time-lag synchronization among cortical areas. *Nature*, 385, 157-161.

Shastri, L. (1999). Advances in SHUTRI – a neurally motivated model of relational knowledge representation and rapid inference using temporal synchrony. *Applied Intelligence*, 63, 69-142.

Shastri, L. & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: a connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences*, 16, 417-494.

Sougné, J.P. (2000). Simulating conditional reasoning containing negations: A computer model and human data. In L.R. Gleitman, & A.K. Joshi. (Eds.) *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society (*pp. 918-923). Mahwah,NJ: Erlbaum.

Sporns, O., Tononi, G., & Edelman, G.M. (1991). Modeling perceptual grouping and figure-ground segregation by means of active reentrant connections. *Proceedings of the National Academy of Science USA*, 88, 129-133.

Steinmetz, P. N., Roy, A., Fitzgerald, P. J., Hsiao, S. S., Johnson, K. O., & Neibur, E. (2000). Attention modulates synchronized neuronal firing in primate somatosensory cortex. *Nature*, 404, 187-189.

Appendix

Equation (5):

$$\sum_{t_c \in F_c} \eta(t - t_c) \approx -\sum_{a \in \Gamma_A} \sum_{t_a \in F_a} w_{c,a} \varepsilon_{c,a}(t - t_a)$$

$$= -\sum_{a \in \Gamma_A} \sum_{t_a \in F_a} w_{c,a} \left(\left[\exp\left(-\frac{t - t_a - \Delta_{c,a}}{\tau_m} \right) - \exp\left(-\frac{t - t_a - \Delta_{c,a}}{\tau_s} \right) \right] H(t - t_a - \Delta_{c,a}) \right)$$

Equation (6):

$$\eta(t) \approx -\sum_{a \in \Gamma_A} w_{c,a} \left(\left[\exp\left(-\frac{t - t_a}{\tau_m} \right) - \exp\left(-\frac{t - t_a}{\tau_s} \right) \right] H(t - t_a) \right)$$

The Conscious-Subconscious Interface: An Emerging Metaphor in HCI

Aryn A. Pyke (apyke@ccs.carleton.ca)
Department of Cognitive Science, Carleton University
Ottawa, K1S 5B6 Canada

Robert L. West (robert_west@carleton.ca)
Department of Psychology, Carleton University
Ottawa, K1S 5B6 Canada

Abstract

Although there already exist traditional metaphors influencing human-computer interaction (HCI) such as the human-human interaction metaphor, some applications emerging in areas from intelligent agents to wearable and context-aware computing have prompted the authors to identify a new metaphor implicitly emerging in HCI. This will be referred to as the conscious-subconscious (C-S) metaphor. The explicit elucidation of this C-S metaphor poises HCI to speak to and leverage research in philosophy and cognitive science pertaining to consciousness and cognition.

Introduction

Although the characterization of consciousness and cognition, per se, are not the primary objectives of the applied field of human-computer interaction (HCI), the authors contend that emerging endeavours in HCI are poising it to speak, perhaps more directly than in the past, to such issues.

To address novel challenges as HCI applications have expanded and diverged, the field has developed a growing toolbox of metaphors to inform and interpret interface design. Two such, traditionally prevalent, metaphors for human-computer interaction are the human-tool interaction metaphor and human-human interaction metaphor (see Marchionini (1995) for a general overview). However, some of the applications in relatively new areas such as wearable, context-aware computing, and intelligent/autonomous agents seem to be promoting the implicit development of a new metaphor which will be referred to as the conscious-subconscious (C-S) metaphor.

In HCI, it is natural to partition computer activity into two classes: interface processes and underlying computational processes which are predominantly opaque to the interface/user. Just as much of the computation that goes on in a computer is not evident in any way through, or even to, the interface, so too, much of the computation that transpires in the mind is opaque; even to the individual's own introspection. This is evident in such activities as throwing a baseball, during which there are no conscious calculations of trajectories.

From this perspective, humans can be modelled as cognitively modular with a conscious process running concurrently with one or more subconscious processes. In some cases our conscious mind is completely unaware of the underlying computation, in other cases it is privy to the product or result of the computation, and sometimes it seems actively involved in the process. In light of it's intermittent and incomplete exposure to inner information, the conscious mind can be viewed in this regard as a form of interface itself.

The following example provides an illustration of how the C-S metaphor can be applied in HCI. The next generation of search engine interfaces will be information retrieval agents (e.g. Lieberman, 1995; Rhodes & Starner, 1996; Stenmark, 1997) which are able to proactively, that is, without an explicit query, search out and bring to the user's attention (consciousness) web documents that might be relevant in the current task context. For example, it might fetch pages on Hamlet while an individual is writing a paper on Hamlet in MS Word. In a similar spirit, subconscious memory processes proactively pop memories into mind (consciousness) that are relevant to the current context, such as when you enter a video store and spontaneously remember that your friend recommended that you should see *The Matrix*. In the sense that such interface agents perform tasks of a similar nature to those performed by human subconscious mechanisms, and exchange information with the conscious facet of the user in a similar manner, such agents could be regarded acting as supplementary or prosthetic subconscious systems.

The explicit elucidation of this C-S metaphor is significant in that, more directly than previous metaphors, it poises HCI to speak to and leverage research in philosophy and cognitive science

pertaining to consciousness and cognition.

A distinctive contribution of HCI in this regard, is that by virtue of it's agenda, it is geared toward actual operationalizations of the concepts and implementations of the processes. Such efforts yield impetus and insight towards identifying and characterising: a) information availability and internal interfacing aspects of conscious experience; and b) the underlying processes necessary to support such experience. It is beyond the scope of this paper to delve into the details of these characterizations beyond a few examples. It is hoped that the identification and systemisation of the metaphor will catalyse interdisciplinary impetus toward an integrated theoretical interpretation and framework.

Human-Computer Interaction: Catering to Consciousness

The flip side of the relevance of HCI to the characterization of consciousness, is the relevance of consciousness to the characterization of HCI. Although in cognitive models of computation, consciousness is often regarded as an aside or epiphenomenon (e.g. Jackendoff, 1987), the user's conscious experience often holds centre stage in the field of human-computer interaction. Familiar conscious experiences during computer use, such as effort, confusion, frustration and impatience are all relevant to the evaluation of an HCI design, and these are the types of experiences that HCI efforts endeavour to minimise.

This focus on phenomenology first is actually shared by an explicit methodological approach to the characterization of consciousness. It is one advocated by William James (1910) who advised that the researcher "begin with the most concrete facts, those with which he has a daily acquaintance in his own inner life".

The C-S Metaphor

Humans, as social animals, are equipped with substantial experience in interacting with each other. Designers can cater to this experience by following a human-human interaction metaphor, and attempt to make interacting with a computer resemble, in some sense, interacting with another individual. In terms of design, the practical implication was that making a better computer (interface) was synonymous with making it appear more human. This pursuit is evidenced in research involved with developing natural language interfaces, endowing computer interfaces with personality trait cues, facial expressions and imbuing them with emotions (e.g. Minsky, 1999; Picard, 1997). As users, we tend to meet the design effort halfway with our inclination to anthropomorphize.

However, some emerging applications in HCI require a phenomenological framework that does not fit well with the human-human, or other traditional interface metaphors. There is a trend for computers to become wearable and equipped with their own sensory systems, (e.g. Rhodes, 1997). They can thus be programmed to be context-aware in terms of both the task and physical environment (Dey & Abowd, 1999, is a good survey). Applications are also being programmed to absorb, and make inferences from, the behavioural patterns of their users. For example, web browsers can keep track of the URL's visited and proactively complete the URL string the user is currently typing by matching it to a habitual one. Extrapolating from the "short hand" ease of interaction enjoyed between close friends due to shared experiences, imagine the potential fluency of interaction with a context-aware, habit-aware, wearable device which could potentially accompany an individual at all times and share all "experiences". In this sense the interaction has the potential not only to be more intimate than ever before in HCI, but also in some respects more seamless than interpersonal interaction, and thus of a different, more continuous and customized character than human-human interaction paradigms could properly inform.

Although such a scenario may not be informed by interpersonal *interaction* paradigms, such an integrated, customized interface has aspects in common with the cognitive *intra*-actions that occur between the conscious and subconscious processes within an individual. Compare the experience of information retrieval from a computer search engine interface, or even by querying another person, with the spontaneous retrieval afforded by internal memory. Memories often pop into mind (consciousness) just as they are needed (Rhodes, 2000), before an explicit query need even be formulated. From the perspective of conscious experience, such retrieval is accomplished by an apparently context-aware unconscious process operating concurrently, autonomously and proactively.

In the human-human interaction metaphor for HCI, the human user is, in principle, regarded somewhat holistically. The computer (interface) is designed to simulate human interface characteristics (e.g. speech recognition, and speech output) by effectively acting as another human with "whom" the user can interact (in the spirit of the Turning Test). In contrast, in the C-S interface metaphor, the computer could be thought to be interacting in a more integrated way, simulating a service style akin to that of the user's *own* subconscious

processing. Several such subconscious systems could be simultaneously active. As with internal subconscious processes, it is not unreasonable to suppose that several could run simultaneously and have some form of intermittent 'interactions' projecting on, or guided from, the explicit stream of conscious experience. Although this is a fairly pre-theoretic model, it is at least consistent with two of the most familiar themes in cognitive science and the philosophy of mind: the modularity of mental processes (e.g. Fodor, 1983) and the introspective opacity of some processes and premises (e.g. Davidson, 1987).

The next section describes intelligent/autonomous interface agents (e.g. Lieberman, 1997) whose characteristics can be applied to C-S metaphor applications.

Intelligent Agents: Aspects of Autonomy and Implicit Interaction

Intelligent/autonomous interface agents (e.g. Liberman, 1997) are software processes/programs that can operate in parallel with the user and autonomously impact the interface. They are best described by example (as provided in the next section), but a brief prefacing discussion is appropriate to emphasise their key characteristics in the current context.

They are able in principle to operate without explicit initiation by the user, and/or continually and concurrently with other explicitly interactive activities of the user. By way of contrast, in strictly turn-taking conversational interfaces, such as the familiar traditional search engine interfaces, the user and the interface take turns (inter)acting, and are dormant while awaiting the response of the other. An agent's ability to act proactively sometimes includes the potential to initiate explicit interaction with the user. For example the agent may explicitly inform the user of a something it has detected in the course of it's independent (though often context-aware) activities.

Conversational interfaces are characterized by communication which tends to be fully explicit in both directions. For intelligent interface agents, there can be explicit and implicit aspects of interaction and information gathering. Though the user may or may not be aware of the agent's activities at any given moment, an agent can be programmed to attend to the user's activities, and base it's activity on such implicitly gleaned information, rather then requiring explicit instruction.

Interface agents can have a broad range of capabilities and characteristics. As such, they are not at all tightly or necessarily coupled to the C-S metaphor. An agent could perform autonomous activity but also employ a conversational interaction style, and fit under the auspices of the human-human interface metaphor. The relevance of such agents here lies in the fact their nature *permits* them to be tailored, if desired, towards applications which do fit with the C-S metaphor. For example, they can run concurrently and invisibly in the background, implicitly ingest information, adapt to the individual, interject with task-relevant information or initiate interruptions.

Analogies between Autonomous Interface Agents and Internal Subconscious Processes

The proposal of a C-S interface metaphor for HCI has been justified by the commonalties noted between the attributes of some such interface agents and those of subconscious processes. The parallels are best exemplified in agents which actually perform similar specific functions to those provided by our own subconscious system. Examples of such agents which are likely familiar to the reader include: (1) the information retrieval agents already alluded to; (2) automatic "typo" correction 'agents' as in MS Word; and (3) automatic string completion 'agents' such as provided in browsers for automatic URL string completion.

(1) The C-S interface metaphor is particularly well suited for information retrieval (IR) applications. For the most part, these electronic IR endeavours have been guided by paradigms from the library sciences. However, a very efficient, personally customised, information retrieval system - our own memory system - provides a nearer and dearer paradigm. It is a testament to the appropriateness of the C-S metaphor that when we forget something, we often assert that "it'll come to us", implying the involvement of an agent/process other than our(conscious)selves.

There are several other phenomena in memory which speak to the conscious/unconscious interaction issue. Most of our memories are subconscious most of the time. Otherwise we would be constantly actively engaged in experiencing all our previous memories (simultaneously)! It is a common experience to have an old memory resurface (not consciously bidden) when returning to an old haunt. It is also common not to be able to remember something, such as the location of one's car keys, despite the fact that one consciously wants to. Such experiences serve to remind us that we do not remember by conscious will alone, but rather through a collaboration (interaction) of the conscious and subconscious mediated by context.

Letizia (Lieberman, 1995) is an example of an autonomous interface agent for web search (also Rhodes, 2000; Stenmark, 1997). Letizia operates in

parallel with the user's browsing activity and is always active, sifting the web space that is "nearby" (linked to) the user's current page of focus. Letizia implicitly gleans the user's current web location by monitoring the user's explicit interaction with the browser, and conducts empirical observation of the user's past and present browsing behaviour to infer aspects of interest. The user profile can be saved, so knowledge persists and accumulates across sessions. This information is used to make relevancy judgements without (prior to) explicit presentation of information to the user. Based on this implicitly acquired information, Letizia proactively displays pages in a separate right frame of the browser that it judges might be of interest. Note that Letizia's search is context driven, such that whenever the user switches pages, Leteiza's context is refocused on the new page. In terms of the user's conscious participation in the interaction, the user is free to ignore or pursue the suggestions shown subtly in a window at the edge of the screen.

(2) Word processing applications such as MS-Word can now be set to automatically correct spelling mistakes as the user types[1]. The operation is subtle and unobtrusive, and in composing this document, this author has remained largely unaware of this corrective activity (though not due to a lack of 'typos').

In a similar vein, human musicians have been observed to automatically internally correct 'musical typos' when playing off sheet music without even being consciously aware of it (Jackendoff, 1987). In both cases, there is a concurrent automatic process that is functioning like a proactive proof-reader.

(3) Automatic string completion 'agents' exemplify other HCI activities that could be interpreted as serving as supplementary subconscious processes. Browsers now can be set to automatically/proactively attempt to complete the URL as the user types. Similarly, in the UNIX command-line interface shell "bash", users can type the first few letters of a filename or command and press tab to invoke the completion process (a beep will sound if there more than one possible match requiring the user to be more explicit). In the UNIX case the user has to explicitly invoke the completion process by pressing tab. In some respects, these examples could be interpreted according to the human-human interface metaphor. Conversing individuals often preemptively complete each other's sentences, and this could be considered an appropriate analogy for automatic string completion. However there are significant reasons why a C-S metaphor could be ultimately considered more appropriate. In principle, even in conversation, it is the speaker herself, not the other individual, who best knows what the correct completion will be. Thus, an intra-individual (subconscious) metaphor rather than an inter-individual metaphor is appropriate for achieving optimal performance. Furthermore, in practise, ultimately wearable computers will be privy to context and user history information on a scale that will make them more 'acquainted' with the user, in their head so to speak, than any human-human acquaintance protocol could be properly equipped to model.

[1] For example, automatically changing adn to and.

Implications for the Characterization of Consciousness and Cognition

A perspective on consciousness which revolves around the human conscious/subconscious (C-S) interface has been taken. The authors have proposed that a C-S metaphor is an emerging (though perhaps implicit) metaphor guiding HCI design. In the course of the discussion it was demonstrated that the operation of several existing autonomous interface agents, such as the type providing automatic URL string completion, can be readily interpreted according to this metaphor. However, the actual development of the existing applications was largely on a piecemeal basis and involved implicit or pre-theoretic application of the metaphor. To the authors knowledge, the C-S metaphor has not been labelled as such nor applied in a systematic manner. It lays the foundation for observing that the different disciplines of philosophy and HCI have converged to concentrate on some of the same issues, and thus are poised to mutually inform/influence each other.

There are two types of challenges inherent in implementations according to the C-S metaphor. First, it is necessary to determine and describe the exact nature of the conscious experience the interface is supposed to engender. This might include such factors as the degrees of implicit and explicit exchange, timing issues and information availability, format and salience. Then it is required to characterize and operationalize underlying computational processes to afford the desired interaction and information exchange. These two challenges are discussed in the following two sections.

Characterizing the Conscious Experience

In the human-human metaphor, much of the interaction is typically conversational and of a very explicit nature. That is, it involves direct, deliberate "communication" with the computer. The user has a conscious experience of ongoing active involvement in the interaction. This includes not only awareness of the information exchanged itself but also

(phenomenology of) intentional initiation and interpretation of exchanges. In contrast, for applications in the C-S style, the trend is towards less (experience of) explicit effort, and greater emphasis on fine tuning on the phenomenological aspects of the information availability.

The HCI endeavour of making interacting with a computer more like interaction with one's subconscious processes provides pressure to make the phenomenology of such interactions more explicit. That is, what are the various types and features of phenomenological projections of unconscious computation processes?

Phenomenology is often exemplified by reference to qualia of sensory experiences such as pain. Considerations regarding the phenomenology of interacting with our subconscious process or computers promote the characterization of a different family/modality of phenomenal experiences. These are the less sensory-centric, conscious correlates of computation and internal communication. For example, there is something it is like to consciously, with effort, conduct some reasoning, or to generate and receive explicit communication.

Some such characterizations, especially in terms of memory, have already preceded the HCI efforts. For example, meta-memory models (Kihlstrom, 1987), the feeling of knowing[2] (Hart, 1965), and tip of the tongue experiences (e.g. James and Burke, 2000). Endeavours in HCI according to the C-S metaphor may catalyse more full characterizations and/or operationalizations of such phenomenology.

For example, with regards to information agents, there are several relevant aspects of memory phenomenology at the level being imitated by the agent. It is necessary that the information become accessible to conscious reasoning and declaration, which falls under the auspices of Block's (1995) access consciousness. Another characteristic is that the time course of it's arrival with respect to changing context be such that it ideally arrives when it is relevant and before the user has a conscious sense of missing or wishing for it. This is a motivating premise of Rhodes' (2000) Just-In-Time information retrieval system. Also it is intended to be rendered accessible in such a way that it enters discretely onto the Cartesian stage without upstaging the focus of the current train of thought. The user is free to disregard or ignore the information. Lieberman (1997) points out the significance of this spontaneous subtle suggestion system in Letizia which avoids having the user make the "context switch" required by conversational interfaces from browsing the space of web pages to explicitly interacting with the search agent.

In some respects interaction with C-S style interface agents cannot possibly be experienced exactly like interaction with internal processes. The information from the interface agent starts out on the screen and is subject to absorption via our sensory system. But what is noteworthy is that some of the convenient and customised character is achieved nonetheless. When viewing memory phenomenology at this level, what seems to matter is that the right data be brought to mind at the right time regardless of whether it was a subconscious or sensory delivery channel.

While such input modality issues might not be brought to the fore in operationalizing the semantic access of declarative memory, such issues might become relevant in postulated cases of agents acting as artificial sensory systems. Consideration of attributes and implications of such artificial sensory agents might shed light on some important phenomenal issues and delineations. It provides a good connection point between HCI inspired insights and existing theoretical frameworks about consciousness such as the one proposed by Block (1995).

Characterizing the Underlying Computation

Determining the character and components of the desired conscious experience, which was discussed in the pervious section, might be far less than half the battle. When it comes to cognition, the portion which projects onto conscious is just the tip of the iceberg. Having established what the desired experience is like, the question becomes: how to make it like that? In order to serve as the subconscious does, it becomes necessary to operationalize the attributes and information resources of the underlying process, at least at a functional level. HCI has faithfully followed this trail far from it's initial focus on the conscious experience. Along the way, it (though perhaps inadvertently) produced computational models that act as unconscious computational processes.

The details are beyond the scope of this paper, but the crux of the matter amounts to characterizing the catch-all notion of context. A preliminary framework within HCI for such a characterization is provided by Dey and Abowd (1999). When it comes to memory, we are unaware of what contextual cues our subconscious process might be using to prompt recall. In HCI mediated memory, the nature of the contextual cues in the computational process need not be identical

[2] The feeling of knowing refers to a scenario in which, although the individual is experiencing difficulty retrieving a memory, they nonetheless feel that it is in there (their subconscious) somewhere.

to the ones leveraged by our subconscious, provided that from a functional perspective they are sufficiently correlated. For example, content stored on a computer can be tagged with time of day and date. Unconscious processes may involve some form of temporal tagging of memories, but not in the same format.

Conclusions

The authors have identified the emergence of an implicit C-S metaphor in HCI, which seems more appropriate to interpret and inform certain context-aware and autonomous agent applications than the traditional human-human interface metaphor. The C-S metaphor was inspired by the observation that, from the perspective of the conscious human experience, autonomous interface agents (e.g. Lieberman, 1997) often have attributes in common with subconscious processes. Most notably they exhibit autonomy, and the capacity for implicit interaction. Despite the existence of various applications that fit the bill, to the authors' knowledge the C-S metaphor has not been systematically identified and exploited. Remembrance Agents (Rhodes & Starner, 1996; Rhodes, 1997) were the closest encountered approximation in this regard.

In the application of the human-human metaphor to HCI, fairly operationalized models already existed on the nature of human-human interaction itself. This is not so much the case for human C-S interactions. This situation provides an opportunity for HCI and cognitive science to mutually inform and give impetus to each other on the C-S interface issue.

Probing the C-S metaphor fosters greater appreciation and awareness of the contribution and role of non-conscious processes in cognition. Newton's quote in homage to his predecessors can be aptly adapted to salute the support structure provided by actual and simulated subconscious processes.

If I have seen farther, it is because I have stood on the shoulders of giants.

-- **Sir Isaac Newton**

References

Block, N. (1995). On a confusion about a function of consciousness. From *Behavioural and Brain Sciences*, 18, 227-247.

Dey, A. K. & Abowd, G.D. (1999). Towards a better understanding of context and context-awareness. *GVU Technical Report GIT-GVU-99-22*, College of Computing, Georgia Institute of Technology.

Fodor, J. (1983). *Modularity of Mind.* Cambridge: MA, MIT Press.

Hart, J. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, 58, 193-197.

James, W. (1910). The Stream of Consciousness. In *Psychology,* Chap. XI. New York: Henry Holt and Co.

James, L. E. & Burke, D. M. (2000). Tip of the tongue, phonological priming and aging. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 26(6), 1378-1391.

Jackendoff, R. (1987*). Consciousness and Computation*. Cambridge, MA: MIT Press.

Kihlstrom, J. F. (1987). The Cognitive Unconscious, *Science*, 237, 1445-1452.

Lieberman, H. (1997). Autonomous Interface Agents. *Proceedings of the ACM conference on computers and human interface*, [CHI-97], New York, NY: ACM Press.

Lieberman, H. (1995). Letizia: An Agent that assists web browsing. *International Joint Conference on Artificial Intelligence*, August 1995. Montreal: QE.

Minsky, M. (1999). 'The emotion machine' from pain to suffering. *Proceedings of the ACM Conference on Creativity and Cognition*, New York, NY: ACM Press.

Marchionini, G. (1995). *Information seeking in electronic environments.* Cambridge, England: Cambridge University Press.

Nelson, T. O. & Narens, L. (1990). Metamemory: a theoretical framework and new findings. *The Psychology of Learning and Motivation*, 26, 125-141.

Picard, R. W. (1997). *Affective Computing*. Cambridge MA: the MIT Press.

Rhodes, B. & Starner, T. (1996). The Remembrance Agent. *AAAI Symposium on Acquisition, Learning and Demonstrations*, Menlo Park, CA: AAAI Press.

Rhodes, B. (1997). The wearable remembrance agent: a system for augmented memory. *Personal Technologies Journal Special Issue on Wearable Computing*, 1, 123-128.

Rhodes, B. (2000). *Just-In-Time information retrieval*. Doctoral Thesis, MIT, Cambridge, MA.

Stenmark, D. (1997). *To search is great, to find is greater: a study of visualization tools for the web.* Unpublished Manuscript. Retrieved October 12, 2000, from the World Wide Web: http://w3.informatik.gu.se/~dixi/publics.htm

Cognitive Uncertainty in Syllogistic Reasoning: An Alternative Mental Models Theory

Jeremy D. Quayle (j.d.quayle@derby.ac.uk)
Cognitive and Behavioral Sciences Research Group,
Institute of Behavioural Sciences, University of Derby,
Mickleover, Derby, DE3 5GX, UK

Linden J. Ball (l.ball@lancaster.ac.uk)
Psychology Department, Lancaster University,
Lancaster, LA1 4YF, UK

Abstract

In this paper we propose a mental models theory of syllogistic reasoning which does not incorporate a falsification procedure and clearly specifies which conclusions will be generated and in what order of preference. It is assumed the models constructed vary in terms of the number of uncertain representations of end terms, and the directness of the representation of the subjects of valid conclusions. These key factors determine which quantified conclusion will be generated, as well as the varying tendency to respond that "nothing follows". The theory is shown to provide a close fit to meta-analysis data derived from past experiments.

Introduction

The categorical syllogism is a deductive reasoning problem comprising two premises and a conclusion (see example below).

> Some artists are beekeepers
> All beekeepers are carpenters
> Therefore, some artists are carpenters

The premises feature three terms which refer to classes of items or individuals: an end term in each premise, and a middle term which appears in both premises. The formal structure of a syllogism is determined by its mood and its figure. The term mood refers to the different combinations of quantifer that can be featured in the premises and conclusion. Four standard quantifiers are used in English language syllogisms: All, Some, No/None, or Some... are not. The term figure refers to the four possible arrangements of the terms within the premises: A-B, B-C; A-B, C-B; B-A, C-B; and B-A, B-C (where A refers to the end-term in the first premise, C refers to the end-term in the second premise, and B refers to the middle term). As each premise can contain one of four quantifiers, and there are four figures, 64 standard premise pairs are possible. The logically valid conclusion to a syllogism is a statement which describes the relationship between the two end terms in a way that is necessarily true, given that the premises are true.

The principal challenges for a theory of syllogistic reasoning are: (1) to explain how people are able to reach the right conclusion for the right reason (i.e., logical competence), and (2) to explain the systematic variations in difficulty and responding between different forms of syllogism (i.e., performance). Responses to syllogisms vary both in terms of the forms of quantified conclusion generated, and in the tendency to respond that there is no valid conclusion. One theory that provides a good fit for the data, and that has received considerable support and attention in the literature is the mental models theory (e.g., Johnson-Laird & Byrne, 1991).

This theory, which can be said to have its routes in early Euler circles, set-based accounts (e.g., Erickson, 1974, 1978), is one of the most comprehensive theories of syllogistic reasoning competence and performance. It assumes that the reasoning process begins with the construction of a mental model of the premises within working memory that makes explicit the minimum amount of information. From this initial model a parsimonious conclusions is formulated, the validity of which is tested in a search for counter-examples– a process which may involve 'fleshing out' the initial model. If a falsifying model cannot be constructed, then the conclusion is generated (since it must be valid), otherwise it is rejected. Reasoners who experience difficulty whilst reasoning will be inclined either to respond with an incorrect quantified response that is consistent with current models, or to say that there is no valid conclusion.

A central assumption of the mental models theory is that syllogisms vary in terms of the number of mental models it is necessary to construct in order to test putative conclusions. With some syllogisms a single model is sufficient for a valid conclusion to be generated (termed one-model syllogisms), with others it is necessary to construct two or three models (termed multiple-model syllogisms). Multiple-model syllogisms

place greater, and less manageable demands on limited working memory resources than one-model syllogisms, and consequently, yield the smallest proportion of valid conclusions and the largest proportion of erroneous no valid conclusion (NVC) responses.

Although good support has been found for mental models accounts of performance (e.g., Johnson-Laird & Bara, 1984), the theory does have some notable weaknesses. First, the theory does not clearly specify how conclusions are formulated, and so cannot adequately explain the clear preference for some conclusions over others. For example, according to Johnson-Laird and Steedman (1978), the theory predicts an average of 3.3 different responses per syllogism. Second, there is increasing evidence to suggest that reasoners construct only a single mental model and do not engage in a falsification process at all (e.g., Polk & Newell, 1995; Bucciarelli & Johnson-Laird, 1999; Newstead, Handley & Buck, 1999).

In spite of these weaknesses, we find many assumptions of the mental models theory to be intuitively plausible, and believe that there is scope for a new Euler-circle inspired mental models account which retains some of these assumptions, specifies the manner in which conclusions are generated more precisely, and incorporates the idea that conclusions are generated after the construction of a single minimal mental representation.

A Cognitive Uncertainty Theory

In accordance with the mental models theory, we propose a theory that assumes people reason with syllogisms by constructing abstract analogical models of the logical relationships between the terms described in the premises within working memory. We would argue that the goal of reasoners when constructing models of syllogisms is to represent mentally both the semantic meaning of each premise and the order of the terms within each premise in the simplest and most probable form. Simple models are those which do not explicitly represent all of the different possible relationships between the end terms. These are constructed for reasons of cognitive economy, since they should not place such high loads on limited working memory capacity as would the consideration of more complex alternative models. Probable models are those which represent the most likely or 'available' situation where a number of alternatives are possible.

Just as the written and spoken forms of terms within premises are read or heard in a particular serial order, so the mental representations constructed of terms within premises are intended to be scanned mentally in a direction which corresponds with these forms of presentation. For ease of explanation we shall refer to this intended scanning direction as 'left to right'.

The Representation of Quantifiers

If a 'universal' quantifier (all or no/none) precedes a term, then the complete class of items or individuals to which the term refers will be represented in the model. For example, in our notation, "All A" would be represented as: [A]

When a premise has the universal, affirmative quantifier all (e.g., "All A are B"), we suggest that reasoners represent the premise as: [A B]

This representation (and the representations constructed for all other forms of premise) is intended to represent both the meaning of the premise and the order of the terms within the premise. It features a description of the class A (the subject of the premise) in terms of the class B. This is the simplest most economical way of representing this premise, since it avoids the need to represent members of the B class who are not also members of the A class. The representation is equivalent to the conditional statement "If it is an A, then it is a B". Just as the statement "All A are B" is intended to be read from left to right (or spoken in a corresponding serial order), so this representation is intended to be scanned from left to right. When scanned from left to right, this representation is unambiguous. However, when scanned from right to left, the representation is ambiguous, suggesting a fallacious identity interpretation of the premise (i.e., ambiguity in the representation leads to the assumption that "All A are B" also means "All B are A").

Twenty eight syllogisms have at least one premise featuring the All quantifier. Eighteen of these are determinate syllogisms. For fourteen of these, the ambiguity of this representation does not affect valid conclusion generation. However, the assumption that participants construct ambiguous representations of All premises can account for all preferred conclusions for the remaining fourteen syllogisms.

The premise "No A are B" (featuring the universal, negative quantifier) would be represented as: [A] [B]

If an 'existential' quantifier precedes a term, then an incomplete class of items or individuals to which the term refers will be represented in the model. For example, "Some A" would be represented as: A) or (A

Grice's (1975) maxim of quantity states that speakers should be as informative as possible, and should not deliberately withhold information which they know to be true. It follows from this notion that it would be wrong for speakers to use the word "some" when they know "all" to be true (see also Begg, 1987; Newstead & Griggs, 1983). We argue, therefore, that although a logician's definition of the quantifier some is "at least one and possibly all", complete classes of items or individuals are not represented when a term is preceded by "some". Hence, we suggest that the premise "Some

A are B" (featuring the existential, affirmative quantifier) would be represented as: [A)B]

With this premise, the possibility that there could be A's that are not B's is not represented in the model, although this may be understood implicitly.

The premise "Some A are not B" (featuring the existential, negative quantifier) would be represented as: (A [B]

In this instance, the possibility that there could be A's that are B's is not represented in the model, but again, this may be understood implicitly.

Model Construction

It is suggested that the construction of models occurs as follows:

1) One of the two premises is picked to be the first premise represented in the model.
2) A model of the first premise is constructed in which the first term in the premise is described in relation to the second term.
3) The model of the first premise is augmented so that it features a representation of the first term in the second premise described in relation to the second term in the second premise. When a universal set is represented in the model of the first premise, and the term referring to that set is preceded by some in the second premise, the universal representation is reduced to an existential representation in the combined model (e.g., syllogisms 3 and 5). As with the first premise, the model is constructed such that the semantic meaning of the quantifier and the serial order of terms in the second premise are retained within the model (see examples below).

1.	2.	3.
Some A are B	Some A are B	No A are B
All B are C	No C are B	Some B are C
[A)B C]	[c] [A)B]	[a] [B)c]
Some A are C	No C are A	No A are C
(valid)	(invalid)	(invalid)
4.	**5.**	**6.**
Some A are not B	No B are A	All B are A
All C are B	Some B are C	No B are C
(A [c B]	[B)c] [a]	[B A] [C]
Some A are not C	Some C are not A	No A are C
(valid)	(valid)	(invalid)

Syllogism 1 exemplifies a situation in which it is necessary to represent members of one end-term class who are also members of the other end-term class in order to construct a model in which the semantic meaning of each premise and the order of the terms within each premise is represented. In contrast, syllogisms 2 to 6 exemplify situations in which it is not necessary to represent members of one end-term class who are also members of the other end-term class.

We acknowledge that the model shown for syllogism 5 is not the only possible model that could be constructed. For example, the possible intersection between the A term and C term, or the possible containment of the A term within the C term could be represented in a model. We would suggest, however, that if it is not necessary to represent an overlap in class membership in a model in order to represent the meaning of each premise and the order of the terms, then no overlap will be represented. The model for syllogism 5 shows no overlap between the A term and the C term, and so, the relationship between these terms shown in the model is equivalent to "No C are A" or "No A are C". This may be considered a rational approach to model construction, since the situation "No X are Y" (where X and Y are two properties picked at random) is almost always true (cf. Chater & Oaksford, 1999). Hence, the model shown for syllogism 5 represents the most probable situation out of a number of possible alternatives.

Certainty and Uncertainty within Models

In some instances, the representation of one end term in relation to the other end term in a model is 'certain'. By this we mean that the class represented by one end term (e.g., the A's in syllogism 4) cannot incorporate members of the class represented by the other end term (e.g., the C's in syllogism 4) unless members of one end-term class (e.g., the A's in syllogism 1) are already members of the other end-term class (e.g., the C's in syllogism 1). In other instances, however (e.g., in syllogisms 2, 3 4 and 5), the representation of one end term is 'uncertain' in relation to the other end term. That is, members of one end-term class who are not represented as being members of the other end-term class could possibly be members of that other end-term class– in our notation, lowercase letters denote uncertain representation of a term. For example, in the model for syllogism 2, as the possibility that there could be some A's that are not B's is not explicitly represented, it is possible that some or all of the C's represented could be A's– hence, the C term is represented by a lowercase letter. In the models for syllogisms 3 and 5, the only certain representations are of the B's that are C's. As some or all of the A's that are represented could be C's, and some or all of the C's that are not B's could be A's, both the A's and the C's are shown in lowercase.

Suggested Conclusion Generation

The conclusion that is initially 'suggested' by a model is the one that follows the serial order in which the terms are represented (left-to-right in our notation). The subject of this conclusion will be the end term on the left. The quantifier that is chosen is determined by the representation of that term. If an existential set is represented, then the conclusion will have an existential quantifier (either some or some... are not), otherwise the conclusion will have a universal quantifier (either All or No). If the first end term in the model is represented outside the second end term, then the conclusion will be negative (either No or Some... are not), otherwise it will be affirmative (either All or Some).

Although conclusions that are suggested by a model will be the preferred responses, the initial conclusion suggested by left-to-right scanning is not always the one that is generated. The initial conclusions suggested by left-to-right scanning of the models constructed for the six example syllogisms are shown beneath the models.

With some syllogisms the representation of the subject of the conclusion suggested by left-to-right scanning is uncertain (e.g., the C term in syllogism 2 and the A term in syllogism 3). Uncertainty of this nature in a model should cause reasoners to lack confidence or certainty over the validity of an initial conclusion, and should motivate them to scan the model from right-to-left in search of an alternative conclusion. If the subject of the conclusion suggested by right-to-left scanning has a certain representation, then this conclusion will be favoured over the conclusion initially suggested (when scanned from right to left the model for syllogism 2 suggests the valid conclusion "Some A are not C" and the model for syllogism 3 suggests the valid conclusion "Some C are not A"). Since two stages of conclusion generation are required with these problems, they load cognitive resources more heavily than those with models where the subject of the conclusion suggested by left-to-right scanning is certain (e.g., syllogism 1). Consequently, they should yield more logical errors– in particular, the generation of invalid conclusions that are consistent with left-to-right scanning (which should be the second most common quantified response)– lower feelings of certainty in participants, and high levels of fallacious NVC responding.

With some indeterminate syllogisms, the subjects of conclusions suggested by both left-to-right and right-to-left scanning have uncertain representations (see syllogisms 8 to 12). With these problems we suggest that the conclusion suggested by left-to-right scanning will be generated more frequently than the one suggested by right-to-left scanning. This is because some reasoners will fail to generate a conclusion through right-to-left scanning due to the extra cognitive demand involved, while most reasoners will be able to generate an initial conclusion through left-to-right scanning. These syllogisms should yield lower feelings of certainty in participants and more NVC responses than syllogisms with models containing one or no uncertain representations.

Direct and indirect representation of the subject With most determinate syllogisms the subject of the valid conclusion is represented 'directly' in the model by members of the class referred to by the subject of this conclusion. For example, the subject of the valid conclusion to syllogism 2 is "Some A", and the representation of this in the model is also "Some A". With some determinate syllogisms, however, the subject of the valid conclusion is represented 'indirectly' in the model by members of the middle-term class who are also members of the class referred to by the subject of this conclusion. For example, the subject of the valid conclusion to syllogism 3 is "Some C", although the representation of this in the model is "Some B". In this instance it is necessary to convert mentally the representation of the subject of the conclusion from "Some B" to "Some C" before a valid conclusion can be generated. As this additional step in the reasoning process is necessary, we suggest that these problems are more difficult than those where the subject is represented directly. Hence, syllogism 3 should yield more logical errors (including a greater proportion of NVC responding) and lower feelings of certainty in participants than syllogism 2.

Implied Conclusion Generation

Not all individuals respond with a conclusion that is suggested directly by a model. A small proportion will respond with a conclusion that is implied by a model. We suggest that implication affects responding in the following way:

- As some and some... are not are given similar interpretations (e.g., see Begg & Harris, 1982), whenever a some conclusion is suggested by a model, a some... are not conclusion will often be generated, and whenever a some... are not conclusion is suggested by a model, a some conclusion will often be generated.
- Conclusions with universal quantifiers (all and no) imply that existential alternatives are also true (some and some... are not respectively).

Categorising Syllogisms

When models are constructed and conclusions generated according to the assumptions we have outlined, five types of determinate syllogism (assigned

labels D1 to D5 in Table 1) and four types of indeterminate syllogism (assigned labels I1 to I4 in Table 1) can be identified. Example syllogisms 1 and 4 are classified as D1 problems, syllogism 6 is a D2 problem, syllogism 2 is a D3 problem, syllogism 5 is a D4 problem, and syllogism 3 is a D5 problem.

Examples of indeterminate syllogisms together with models and suggested conclusions are given below.

7.	8.	9.
All A are B	Some B are not A	Some A are not B
Some B are C	No C are B	Some C are B
[A B) C]	[c] (B [a]	(a [c) B]
Some A are C	NVC	NVC
	No C are A	Some A are not C
10.	11.	12.
No B are A	Some B are not A	Some B are A
No B are C	Some B are C	Some C are B
[B] [a] [c]	(B [a] [B) c]	[B) a] [c) B]
[a] [B] [c]	(a [B) c]	[c) B) a]
or	or	NVC
[c] [B] [a]	((c B [a]	Some C are A
NVC	NVC	
No A are C	Some A are not C	
or	or	
No C are A	Some C are not A	

Example syllogism 7 is an I4 problem, syllogisms 8 and 9 are I1 problems, syllogisms 10 and 11 are I3 problems, and syllogism 12 is an I2 problem. With two of these (types I4 and I1), it is possible to construct a model in which the semantic meaning of each premise and the order of terms in the premises are represented with little difficulty. With the remaining two, however, the construction of a 'unified' model is less straightforward, as either: (1) the premises suggest models with a split representation of the middle term (type I2 and some I3), or (2) the premises do not dictate the order in which the end terms should be represented in a model (type I3). With type I2 and I3 problems, reasoners may feel highly confident that there is no necessary relationship between the end terms, and so, respond that there is NVC. However, as reasoners display a bias towards the generation of quantified responses where NVC responses are appropriate (e.g., Johnson-Laird & Steedman, 1978; Johnson-Laird & Bara, 1984), with I2 and I3 syllogisms we have considered strategies reasoners may adopt in order that models might still be constructed and quantified conclusions generated. With I2 syllogisms a quantified conclusion may be generated if reasoners construct a model in which a representation of the end term which forms the subject of a some premise (e.g., "Some C are B" in syllogism 12) is contained within the class referred to by the middle term in the other premise (see second model for syllogism 12). There are good reasons for assuming that models of this nature are constructed: (1) they are able to represent the semantic meaning of each premise and the order of the terms within each premise, and (2) they feature a representation of a highly probable relationship between the end terms.

With I3 syllogisms, a unified model may be constructed after applying a simple conversion procedure to one of the premises (i.e., by switching the two terms in a premise around without changing the quantifier). Once conversion has taken place, it is possible for reasoners to construct models like those constructed for I1 or I2 syllogisms (depending upon mood) and to generate quantified conclusions (see second models for syllogisms 10 and 11).

Table 1: Nine types of syllogism as a function of figure, together with NVC rankings.

Type	AB BC	BA CB	AB CB	BA BC	Total	NVC Rank
D1	4	4	4	4	16	1
D2	1	1	0	3	5	2
D3	0	0	2	0	2	3
D4	0	0	0	2	2	3
D5	1	1	0	0	2	4
I1	5	5	2	0	12	5
I2	2	2	0	0	4	6
I3	0	0	5	7	12	6
I4	3	3	3	0	9	2.5

Rank values have been assigned to each type of syllogism in Table 1 according to the level of NVC responding that should be associated with them according to the theory (a 1 ranking denotes a low level of NVC responding). With the exception of I2 and I3 syllogisms (where high levels of NVC responding should be associated with strong feelings of certainty), these rankings should correlate negatively with the levels of certainty associated with each type of syllogism.

Predicting performance

We have used Chater and Oaksford's (1999) meta-analysis data (on conclusion quantification and NVC responding) together with Johnson-Laird and Steedman (1978, Experiment 1 and 2), and Johnson-Laird and Bara's (1983, Experiment 3) data (on conclusion term order) in order to test how well the cognitive uncertainty theory explains syllogistic reasoning performance.

Quantification

The theory makes ranking predictions for most common conclusions, as well as second, third and fourth most common conclusions. Predicted rankings match data rankings in 175 out of 256 cases. When the proportions of quantified responses associated with correct matches are summed, the theory accounts for nearly 91% of quantified responses in exact order of commonality.

Term Order in Preferred Conclusions

The theory makes a definite prediction about the order of terms in preferred conclusions for 50 of the 64 premise pairs. It matches the term order for preferred conclusions in 47 of these, and therefore, directly predicts the term order for over 73% of preferred quantified responses in the data—although the theory can accommodate over 95% of these.

NVC responding

Rank values have been assigned to each syllogism based on predicted feelings of certainty in participants, and how much NVC responding would be expected according to the theory (see Table 1). There is a significant correlation between these rankings and the percentages of NVC responses in the meta-analysis data (Spearman's Rho = .885, $p < .001$). The theory, therefore, accounts for 78% of the variance in the NVC data.

Discussion

A new theory of syllogistic reasoning, inspired by early Euler circles, set-based, explanations has been proposed. The theory has been shown to provide a good fit to meta-analysis data derived from past experiments, in terms of: (1) predicting order of preference for quantified responses, (2) predicting term-order in preferred conclusions, and (3) accounting for the varying tendency to give NVC responses.

The new theory was developed in response to an increasing awareness that key assumptions of the mental models account proposed by Johnson-Laird and colleagues were failing to receive support in the literature. In particular, serious doubts have been raised concerning the conclusion falsification and fleshing out processes outlined in the mental models theory. The new account overcomes this problem by suggesting that only a single model is constructed based on an interpretation of the semantic meanings of the premises. Models may be logically accurate, or suggest a fallacious or ambiguous interpretation of All quantifiers. Difficulty is said not to be caused by cognitive loads associated with model construction, but instead by the cognitive loads associated with 'searching' a model in order to identify a conclusion that does not feature an uncertain representation of the subject.

The mental models account may also be criticised for not adequately specifying the manner in which conclusions are formulated, and hence, not explaining why clear orders of preference for different quantified conclusions are evident in the data (e.g., see Chater & Oaksford's meta-analysis of past experiments). The new theory clearly specifies which conclusion will be identified first from a model, whether other conclusions will be identified that are also suggested by a model, and which conclusions will be generated to a lesser degree after being implied by a model. Importantly, clear psychological justifications are given for the predicted orders of preference for these quantified responses.

References

Begg, I. (1987). Some. Canadian Journal of Psychology, 41, 62–73.

Begg, I., & Harris, G. (1982). On the interpretation of syllogisms. Journal of Verbal Learning and Verbal Behaviour, 21, 595–620.

Bucciarelli, M., & Johnson-Laird, P.N. (1998). Strategies in syllogistic reasoning. Cognitive Science, 23, 247–303.

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. Cognitive Psychology, 38, 191–258.

Erickson, J.R. (1974). A set analysis theory of behaviour in formal syllogistic reasoning tasks. In R.L. Solso (Ed.), Theories of cognitive psychology: The Loyola Symposium. Hillsdale, N.J.: Lawrence Erlbaum Associates Inc.

Erickson, J.R. (1978). Research on syllogistic reasoning. In R. Revlin & R.E. Mayer (Eds.), Human reasoning. Washington, D.C: Winston.

Grice, P. (1975). Logic and converstion. In P. Cole & J.L. Morgan (Eds.), Studies in syntax, Vol. 3: Speech acts. New York: Academic Press.

Johnson-Laird, P.N., & Bara, B.G. (1984). Syllogistic inference. Cognition, 16, 1–62.

Johnson-Laird, P.N., & Byrne, R.M.J. (1991). Deduction. Hove: Lawrence Erlbaum Associates Ltd.

Johnson-Laird, P.N., & Steedman, M. (1978). The psychology of syllogisms. Cognitive Psychology, 10, 64–99.

Newstead, S.E., & Griggs, R.A. (1983). Drawing inferences from quantified statements: A study of the square of opposition. Journal of Verbal Learning and Verbal Behaviour, 22, 535–546.

Newstead, S.E., Handley, S.J., & Buck, E. (1999). Falsifying mental models: Testing the predictions of theories of syllogistic reasoning. Memory and Cognition, 27, 344–354.

Polk, T.A., & Newell, A. (1995). Deduction as verbal reasoning. Psychological Review, 102, 533–566.

Using a Triad Judgment Task to Examine the Effect of Experience on Problem Representation in Statistics

Mitchell Rabinowitz (mrabinowitz@fordham.edu)
Graduate School of Education, Fordham University
New York, NY 10023 USA

Tracy M. Hogan (tehogan@fordham.edu)
Graduate School of Education, Fordham University
New York, NY 10023 USA

Abstract

This research investigated whether the differences found between novices and experts in using surface and deep structures to categorize problems applied to the domain of statistics. Also explored was whether the methodology of a triad judgment task was reliable in discriminating how beginning and advanced students represent statistics problems. The task was designed in which source problems shared either structural features (t-test, correlation, or chi-square) or surface similarity (story narrative) with the target problem. Graduate students (N = 101) with varying levels of experience in the domain of statistics were asked to chose which source problem "goes best" with the target problem for each triad. Students with advanced experience in statistics tended to represent the problems on the basis of deep, structural features while beginning students tended to rely on surface features. Discussion on the effectiveness of the methodology employed and potential uses in other domains is presented.

Introduction

Students learning statistics are required to learn a set of interacting skills. First, they need to become familiar with statistical procedures and how to use them (computing formulas). Second, they need to be able to recognize when to use those statistical procedures. The first set of skills is procedural in nature, i.e., they need to learn formulas and know how to execute the computation (or the statistical packages). The latter type of skill is representational, i.e., they need to be able to perceive and represent features within contexts that suggest which procedures should be used.

Previous research (Adelson, 1981; Chi, Feltovich & Glaser, 1981; Chase & Simon, 1973; Hardiman, Dufresne & Mestre, 1989; Schoenfeld & Herrmann, 1982) has shown that experts and novices within a domain represent problems within that domain on the basis of a different set of features. Bransford, Brown & Cocking (1999) report that this difference, in part, lies in knowledge organization. Expert knowledge centers on core concepts and big ideas found within the domain while novices rely on isolated facts and do not connect these facts in a way that allows them to generate inferences. For example, Chi and colleagues (1981) found that participants with advanced experience in physics sorted problems in their discipline on the basis of structural features, including the laws and principles of physics. When asked to sort the same problems, novices represented, and subsequently sorted the problems on the basis of surface features, such as the object being manipulated in the problem.

Quilici and Mayer (1996) argue that while surface features are generally more salient than structural features for novices, successful analogical transfer is dependent upon the recognition of structural similarities among problems. Consequently, they investigated the role of examples in how students learn to categorize statistic word problems. Their findings suggest that exposure to examples influences inexperienced students' structural schema construction, particularly when the example problems emphasize structural characteristics versus surface characteristics. Quilici and Mayer contend that their study merits further research concerning the conditions under which students rely on surface features or structure features in categorizing problems. In that Quilici and Mayer's participants were limited to those with little or no knowledge about statistics, further research concerning the effect of experience on problem representation is warranted.

This study was designed to replicate the expert/novice difference in perception and representational skill in the context of statistics problems. The purpose of this study was two-fold. First, the study investigated whether the differences found between novices and those with advanced experience in statistics use surface and/or deep structures to categorize problems applied to the domain of statistics. Second, this research explored whether the methodology of a triad judgment task was reliable in discriminating how beginning and advanced students represent statistics problems. Consequently, this study extended Quilici and Mayer's research (1996) by determining if those with advanced training in statistics

do indeed cue in on the structural features of a statistical word problem.

To complete this extension, a triad judgment task was designed and administered to individuals with varying levels of statistical experience. According to Hardiman, Dufresne & Mestre (1989), the triad judgment task offers several advantages over the traditional sorting task used in previous research (Chi et. al, 1981; Quilici & Mayer, 1996). First, participants are able to concentrate on individual problem sets rather than being presented with a stack of cards to sort simultaneously. Second, the task allows for large-group administration and ease in scoring. The design of the triad task in this study was similar in nature to that employed by Hardiman and colleagues' research (1989). However, it differed in that this research examined problem representation in the domain of statistics while theirs was grounded in the field of mechanics.

The judgment task required participants to identify which of two given source problems "goes best" with a target problem (Figure 1). The source problems shared

Target Problem	
Source Problem 1	Source Problem 2

Figure 1
Structure of Triad Problems

either similar surface features or structural features with the target problem. Surface features were similar in that the story narrative shared common characteristics while similar structural features involved the requirement of the same statistical test (t-test, correlation and chi-square). Surface features included similar story characters (personnel expert, meteorologist, college dean and psychologist) and similar dependent/independent variables (words typed per minute/experience of typists, annual rainfall/average yearly temperature, grade point average/reading score, number of errors on a test/amount of sleep). The structural features included the nature of the independent variable (one group or two independent groups) and the nature of the dependent variable (continuous or categorical).

Using the statistics word problems from Quilici and Mayer's study (1996), 18 triads were designed to investigate whether this judgment task would discriminate between those representing the problems using deep, structural features with those relying on surface features. To do this, we administered the task to students with varying levels of experience in the domain of statistics. We hypothesized that students with more advanced statistical experience would predominantly represent problems based on structural features while students with less statistical experience would tend to represent the problems based on surface features.

Method

Participants

The participants were 101 graduate students with a varied amount of experience in statistics. Those with no prior statistics courses totaled 27 participants, 33 participants completed one course, 13 finished two courses, 10 had completed three courses, six participants completed four courses, eight participants finished five courses, three participants had completed six courses and one participant completed eight courses. All individuals who volunteered to participate in this study earned course extra-credit.

Problem Task

A triad judgment task was used to investigate the features that people use to represent common statistics problems. The task involved the presentation of three statistical problem statements- one target problem and two source problems. Participants were asked to read each set and judge which of the two source problems "goes best" with the target problem. Comparisons were based upon two features: surface and structure. Surface features were defined by the narrative characteristics (i.e., "After comparing weather data for the last 50 years, a meteorologist claims...") and structural features were defined by requisite statistical tests (t-test, correlation, chi-square).

There were three sets of comparison types that participants were asked to evaluate (Appendix). In the first comparison, one source problem shared only similar surface features to the target problem while the other source problem shared only similar structural features. Thus, Comparison One was considered Similar Narrative / Dissimilar Structure - Similar Structure / Dissimilar Narrative (SN/DS-SS/DN). In the second comparison, one source problem shared no similarities in either surface or structure while the other shared only similar structure to the target problem. Thus, Comparison Two was considered Dissimilar Narrative / Dissimilar Structure - Similar Structure/Dissimilar Narrative (DN/DS-SS/DN). In the third comparison, one source problem shared only similar surface features to the target problem while the other shared neither surface nor structural similarities. Thus, Comparison Three was considered Similar Narrative / Dissimilar Structure - Dissimilar Structure /

Dissimilar Narrative (SN/DS-DS/DN). Each participant was presented six triads per comparison for a total of 18 triads.

Procedure

Participants were given a packet that contained the 18 triad problems and a cover sheet. On the cover sheet, the participants recorded background information including prior statistics courses, education level, and gender. Participants were tested during class and were given as much time as needed to complete the task.

Scoring

A maximum score of 18 points, at six points per comparison type was possible. Participants scored one point per triad under Comparison One (SN/DS-SS/DN) and Comparison Two (DN/DS-SS/DN) if they selected on the basis of structural features. For Comparison Three (SN/DS-DS/DN), participants scored one point if they selected similar surface features in that neither comparison problems shared structural features with the target problem. Thus, a higher score implies a tendency towards choosing the structural dimension or the surface dimension where appropriate.

Results

A correlation analysis was conducted to examine in greater depth the relationships between the level of experience (as measured by the number of statistics courses an individual completed) and the three comparison types. Findings suggest a significant relationship between number of courses and total score, $r = .39, p < .01$. This suggests that the more experience an individual has in statistics, the more likely they are to make more structural comparisons between two statistical passages.

While there was a significant correlation between the number of courses completed and total score on the triad judgment task, there were differences found among the three comparison types. Specifically, only Comparison One (SN/DS-SS/DN) and Comparison Two (DN/DS-SS/DN) were significantly correlated with the number of courses ($r = .35, p < 01, r = .39, p < .01$, respectively). These results, taken together, suggest that the more experience one has in statistics, the more likely he/she is to group statistical passages according to similar methodologies. As expected, there was no significant correlation between experience level and Comparison Three (SN/DS-DS/DN). If neither of the two source problems shared structural features with the target problem, individuals, regardless of experience, choose upon the basis of surface features.

In addition, to investigate whether individuals with more experience in statistics performed differently on the three comparison types as did novices, a repeated measures ANOVA was conducted. Individuals were grouped into three levels of experience in the domain of statistics. Level One included participants who had taken either zero or one course (n=60), Level Two reflected participants that had completed either two or three courses (n=23) and Level Three included participants that had completed four or more statistics courses (n=18). Means and standard deviations for scores on the three comparison types for each experience level are presented in Table 1. The

Table 1: Means and Standard Deviations for Comparison Type by Experience Level.

Level	n	Type I		Type II		Type III	
		M	SD	M	SD	M	SD
One	60	1.73	1.68	3.87	1.24	4.87	1.32
Two	23	2.13	1.94	4.04	.88	4.52	1.20
Three	18	3.44	1.61	4.94	1.06	4.16	1.50

significant interaction between experience level and comparison type suggests a relationship between the level of experience and the way the individual represents the particular statistical problem, $F(2, 98) = 4.94, p < .01$. Tukey's HSD test indicated that those in level three performed significantly different than those in levels one and two. The significant main effect of experience level indicates that individuals with more training in statistics represent statistical passages in ways that are more expert, $F(2, 98) = 6.67, p < .01$. The significant main effect of comparison type suggests that individuals, regardless of level of experience, do not respond in the same way to the different problems found in the triad judgment task, $F(2, 98) = 44.89, p < .01$.

Discussion

In this study, two questions were tackled. The first question was, How do beginning and advanced students in statistics compare in the way they represent statistical word problems? The analyses revealed several contrasts. It was shown that those with advanced experience tended to look for similar deep structures in the word problems presented within the triads. Conversely, the findings suggest that novices relied more heavily on the surface features to match a source problem with a target problem. However, when presented with comparisons types where neither of the source problems shared deep structural features with the target problem, all students, regardless of experience, selected on the basis of similar surface features.

The second question was, Can a triad judgment task be used to reliably discriminate how beginning and advanced students represent statistics word problems on

either the basis of structural features or surface features? On the basis of earlier research (Chi, Feltovich & Glaser, 1981; Hardiman, Durfresne & Mestre, 1989), we reasoned that those with advanced training in statistics would make selections based on structural features while those with less training would select on the basis of surface features in a triad judgment task. Findings were consistent with our prediction. This suggests that the triad judgment task may indeed be a promising methodology to employ in studies where sorting tasks are traditionally used.

This study yields implications for educators of statistics. First, instruction in statistics should address the nature of problems and their structural components (e.g., type of data presented and the driving question of the problem). Second, learners should be provided with explicit instruction in recognizing similarities of problems based on core concepts, a skill requisite of experts (Bransford, Brown and Cocking, 1999).

This study certainly contributes to the relatively narrow research base of experts-novices in statistics, yet further studies are needed. Specifically, more studies are needed to explore the circumstances that promote the transition from using surface characteristics to deep structural features in representing problems.

References

Adelson, B. (1981). Problem solving and the development of abstract categories in programming languages. Memory & Cognition, 9, 422-433.

Bransford, J.D., Brown, A.L., & Cocking, R.R. (Eds.). (1999). How people learn: Brain, mind, experience, and school. Washington DC: National Academy Press.

Chase, W.G., & Simon, H.A. (1973). Perception in chess. Cognitive Psychology, 4, 55-81.

Chi, M.T.H., Feltovich, P.J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5, 121-152.

Hardiman, P.T., Dufresne, R., & Mestre, J.P. (1989). The relation between problem categorization and problem solving among experts and novices. Memory and Cognition, 17, 627-638.

Quilici, J.L., & Mayer, R.E. (1996). Role of examples in how students learn to categorize statistics word problems. Journal of Educational Psychology, 88, 144-161.

Rabinowitz, M., & Glaser, R. (1985). Cognitive structure and process in highly competent performance. In F.D. Horowitz and M. O'Brien (Eds.), The gifted and talented: A developmental perspective. Washington DC: American Psychological Association.

Shoenfeld, A.H., & Herrmann, D.J. (1982). Problem perception and knowledge structure in expert and novice mathematical problem solvers. Journal of Experimental Psychology: Learning, Memory, & Cognition, 8, 484-494.

Appendix

Comparison One: Similar Narrative/Dissimilar Structure - Similar Structure/Dissimilar Narrative (SN/DS-SS/DN)

Target: After examining weather data for the last 50 years, a meteorologist claims that the annual precipitation is more likely to be above average in years when the temperature is above average than when temperature is below average. For each of the 50 years, she notes whether the annual rainfall is above or below average and whether the temperature is above or below average.

Source 1: After examining weather data for the last 50 years, a meteorologist claims that the annual precipitation varies with the average temperature. For each of 50 years, she notes the annual rainfall and average temperature.

Source 2: A college dean claims that a group of good readers contains more honors students than a group of poor readers. For each of 100 first year college students, a reading comprehension test was used to determine whether the student was a good or poor reader and grade point average (GPA) was used to determine whether or not the student was an honors student.

Comparison Two: Dissimilar Narrative/Dissimilar Structure – Similar Structure/Dissimilar Narrative (DN/DS-SS/DN)

Target: A college dean claims that good readers earn better grades than poor readers. The grade point averages (GPA) are recorded for 50 first-year students who scored high on a reading comprehension test and for 50 first-year students who scored low on a reading comprehension test.

Source 1: A psychologist tests the hypothesis that people who are fatigued also lack mental alertness. An attention test is prepared which requires subjects to sit in front of a blank TV screen and press a response button each time a dot appears on the screen. A total of 110 dots are presented during a 90-minute period, and the psychologist records the number of errors for each subject. Twenty subjects are selected; half are tested after being kept awake for 24 hours and half are tested in the morning after a full nights sleep. Based on the number of errors on their test, each subject is also labeled as high or low in mental alertness.

Source 2: A personnel expert wishes to determine whether experienced typists are able to type faster than inexperienced typists. Twenty experienced typists (i.e., with 5 or more years of experience) and 20

inexperienced typists (i.e., with less than 5 years of experience) are given a typing test. Each typists average number of words typed per minute is recorded.

Comparison Three: Similar Narrative/Dissimilar Structure - Dissimilar Structure/Dissimilar Narrative (SN/DS-DS/DN)

Target: After examining weather data for the last 50 years, a meteorologist claims that the annual precipitation varies with the average temperature. For each of 50 years, she notes the annual rainfall and average temperature.

Source 1: After examining weather data for the last 50 years, a meteorologist claims that the annual precipitation is greater in years with below average temperature than in years with above average temperature. She notes the annual rainfall for each of 25 years that had above average temperatures as well as 25 years that had below average temperatures.

Source 2: A psychologist tests the hypothesis that people who are fatigued also lack mental alertness. An attention test is prepared which requires subjects to sit in front of a blank TV screen and press a response button each time a dot appears on the screen. A total of 110 dots are presented during a 90-minute period, and the psychologist records the number of errors for each subject. Twenty subjects are selected; half are tested after being kept awake for 24 hours and half are tested in the morning after a full nights sleep. Based on the number of errors on their test, each subject is also labeled as high or low in mental alertness.

Perceptual Learning Meets Philosophy: Cognitive Penetrability of Perception and its Philosophical Implications

Athanassios Raftopoulos (raftop@ucy.ac.cy)
Department of Educational Sciences
University of Cyprus
P.O. Box 20537
Nicosia 1678, Cyprus

Abstract

The undermining of the cognitive impenetrability of perception has led to the abolition of the distinction between *seeing* and *seeing as*, clearing the way for the relativistic theories of science and meaning, since perception becomes theory-laden. Hence the existence of a theory-neutral basis, on which a rational choice among alternative theories could be based, is rejected and scientific theories become incommensurable. One of the arguments against the cognitive impenetrability of perception is based on evidence from neuroscientific studies that suggest the plasticity of the visual cortex, in the sense that there can be some local rewiring of the neural circuitry of the early visual system, as a result of experience. This is taken to constitute evidence that the early vision is cognitively penetrable. In this paper I argue that the evidence concerning perceptual learning does not entail the cognitive penetrability of perception. To that end I discuss the issue of perceptual learning and claim that this learning is task and data-driven and not theory-driven. The process is mediated by the allocation of attention, which though cognitively penetrable, allows only an indirect form of cognitive penetrability of perception. In the last part I elaborate on the significance of this indirect penetrability, as opposed to the direct penetrability by cognition, and discuss its implications for the issue of the incommensurability of scientific theories. My conclusion is that attention can be controlled across different theoretical backgrounds, and thus, that the indirect cognitive penetrability does not entail incommensurability.

Introduction

The undermining of the cognitive impenetrability of perception has led to the abolition of the distinction between *seeing* and *seeing as* (Gregory, 1974; Hanson, 1958; Kuhn, 1962), clearing the way for the relativistic theories of science and meaning, since perception becomes theory-laden (what we see depends on our expectations, beliefs, and so forth). Hence the existence of a theory-neutral basis, on which a rational choice among alternative theories could be based, is rejected and scientific theories become incommensurable. There can be no communication between scientists that belong to different scientific paradigms, because there is not a theory-neutral perceptual basis that could resolve matters of meaning. Instead, empirical evidence, becomes part of a paradigm or a theoretical research program, being modulated by its theoretical commitments. Thus, proponents of different paradigms or research programs either perceive different worlds (strong version of relativism; Kuhn, 1962), or cannot compare their theories on the basis of some neutral empirical evidence but must search for other criteria of theory evaluation (medium version of relativism; Churchland, 1989).

One of the arguments against the cognitive impenetrability of perception and in favor of its theory-ladenness is based on evidence from neuroscientific studies that suggest the plasticity of the visual cortex, and more specifically, on evidence that there can be some local rewiring of the neural circuitry of the early visual system, as a result of experience (the phenomenon of perceptual learning).

The plasticity of the brain and the possibility of rewiring of the neural circuitry of the perceptual systems, as a result of acquiring knowledge, goes against the view (Fodor, 1983; Pylyshyn, 1999) that some part of vision, the early vision, is cognitively impenetrable, and shows (Churchland, 1988) that perception is cognitively penetrable, in so far as learning, which is a cognitively driven process, affects even those circuits that are involved in early vision.

In this paper I argue that the evidence concerning perceptual learning does not entail the cognitive penetrability of perception. To that end I discuss the issue of perceptual learning and claim that this learning is task and data-driven and not theory-driven.

Being the former it allows only an indirect form of cognitive penetrability of perception. In the last part I elaborate on the significance of this indirect penetrability, as opposed to the direct penetrability by cognition, and discuss its implications for the issue of the incommensurability of scientific theories. My conclusion is that this indirect penetrability does not entail incommensurability.

I have spoken of perception. This term is not employed consistently in the literature. Sometimes "perception" purports to signify our phenomenological experience, and thus, "is seen as subserving the recognition and identification of objects and events" (Goodale 1995, 175). Since I do not use the terms the same way, I will introduce some terminology.

I call *sensation* all processes that lead to the formation of the retinal image (the retina's photoreceptors register about 120 million pointwise measurements of light intensity). This image, which initially is cognitively useless, is gradually transformed along the visual pathways in increasingly structured representations (such as, edges, boundaries, shapes, colors) that are more convenient for subsequent processing. I call these processes that transform sensation to a representation that can be processed by cognition *perception.* Perception includes both low-level and intermediate-level vision and, I claim, is bottom-up. In Marr's (1982) model of vision the *21/2D sketch* is the final product of perception. As I shall argue next perception is non-epistemic, that is, it is independent of specific-object knowledge. All subsequent visual processes fall within *cognition*, and include both the post-sensory/semantic interface at which the object recognition units intervene as well as purely semantic processes, that lead to the identification of the array (high-level vision). At this level we have observation (Marr's *3D model*), which is a cognitive activity.

Perceptual Learning

There is growing evidence for the diachronic penetrability of perceptual systems and for local rewiring of their neural circuits (Ahhisar and Hochstein, 1993; Antonini, Strycker, and Chapman, 1995; Karni and Sagi, 1995; Stiles, 1995). Is there a way to reconcile the notion of cognitive impenetrability of perceptual systems with this evidence? Fodor thought that there is not any, and that this issue would be resolved with the findings of empirical research. Should empirical research show perceptual learning to be possible, then the encapsulation of his input modules would have been proved false. The evidence suggests that perceptual systems are indeed diachronically, in the long run, open to some rewiring of the patterns of their neural connectivity, as a result of learning. These systems are to some extent plastic. But this does not mean that they are cognitively penetrable. Let us see why.

Research shows that changes can be induced in visual cortical neural patterns in response to learning. More specifically, visual processing at all levels may undergo long-term, experience-dependent changes. The most interesting form of learning is "slow learning", because it is the only type that causes structural changes in the cortex (formation of new patterns of connectivity). Such learning can result in significant performance improvement. For example, one may learn with practice to perform better at visual skills involving target and texture discrimination and target detection, and to learn to identify visual patterns in fragmented residues of whole patterns (priming). Performance in these tasks was thought to be determined by low-level, stimulus-dependent visual processing stages. The improvement in performance in these tasks, thus, suggests that practice may modify the adult visual system, even at the early levels of processing. As Karni and Sagi (1995, 95-6) remark "[L]earning (acquisition) and memory (retention) of visual skills would occur at the earliest level within the visual processing stream where the minimally sufficient neuronal computing capability is available for representing stimulus parameters that are relevant input for the performance of a specific task."

Karni and Sagi (1995) suggest that slow learning is independent of cortico-limbic processing, which is responsible for top-down processes and, through the interaction of the limbic system with the visual pathways, responsible for conscious object recognition. It is also independent of factors involving semantic associations. Yeo, Yonebayashi, and Allman (1995) suggest that priming facilitates the neural mechanisms for processing images and that the cortex can learn to see ambiguous patterns by means of experience-induced changes in functional connectivity of the relevant processing areas. Thus, priming involves a structural modification of basic perceptual modules. Practice with fragmented patterns leads to the formation of the "priming memory" which may be stored in the cortical visual areas. Long-term potentiation (LTP) may be the mechanism implementing these architectural changes by establishing experience-dependent chains of associations and dissociations.

Slow learning-induced architectural modifications are "experience dependent" (Greenough, et al., 1993), in that they are controlled by the "image" formed in the retina. But, although learning and its ensuing functional modifications occur in those neuronal

assemblies that are activated by the retinal image, still some extra-retinal factor should provide the mechanism that will gate functional plasticity. Although many neuronal assemblies are activated by the retinal image, learning occurs only in those assemblies that are behaviorally relevant. This is called the "gating of neuronal plasticity".

The factor that modulates gating is the demands of the task. They determine which physical aspects of the retinal input are relevant, activating the appropriate neurons. Functional restructuring can occur only at these neuronal assemblies. The mechanism that accomplishes this is attention. Focusing attention ensures that the relevant aspects of the input are further processed. Attention intervenes before the perceptual processes; selective attentional shifts to specific parts of the visual field precede saccadic eye movements directed to these parts (Hoffman and Subramaniam, 1995). Attention seems to determine the location at which search will be conducted and/or the relevant features that will be picked-up, since focal attention may enhance the output of the salient feature detectors by lowering firing thresholds (Egeth, *et al*., 1984; Kahneman and Treisman, 1992; McCleod *et. al*., 1991). There is indeed ample evidence for the necessary role of attention in perceptual learning (Ahissar and Hochstein, 1995) and for the role of attention in learning to perceive ambiguous figures (Kawabata, 1986; Peterson and Gibson, 1991).

Recall that slow learning is independent of recognition and semantic associative memory. Most of the priming effects are associated with identification and discrimination of relative spatial relations and extraction of shapes. This brings to mind Hildreth and Ulmann's (1989) *intermediate level* of vision. The processes at this level (the extraction of shape and of spatial relations) are not bottom-up, but do not require the intervention of specific-object knowledge, since they require the spatial analysis of shape and spatial relations among objects. This analysis is task dependent but not theory-driven, that is, it is not directly penetrated by cognition.

I have spoken of "specific-object knowledge" and claimed that this kind of knowledge does not intervene in slow learning, and does not threaten cognitive impenetrability of perception. I would like to explain the qualification "knowledge about specific objects". Even if perception turns out to be bottom-up in character, still it is not insulated from knowledge. Knowledge intrudes on perception, since early vision is informed and constrained by some general world principles that reduce indeterminancies in information. They are general assumptions about the world constraining visual processing (Marr, 1982; Ulmann, 1979). These principles however are not the result of explicit knowledge acquisition about specific objects but are general reliable regularities about the optico-spatial properties of our world.

This knowledge is implicit, in that it is available only for the processing of visual information, whereas explicit knowledge is available for a wide range of cognitive applications. Implicit knowledge cannot be overridden. The general constraints hardwired in the visual system can be overridden only by other similar general constraints with which they happen to compete (although no one knows yet how the system "decides" which constraint to apply). Still, one cannot decide to substitute it with another body of knowledge, even if one knows that under certain conditions this implicit knowledge may lead to errors (as is the case with the visual illusions). This theoretical ladenness, therefore, cannot be used as an argument against the existence of a theory-neutral ground, because perception based on a shared theory is common ground.

Slow learning, thus, takes place under specific retinal input and attention-dependent conditions. Although the allocation of attention is clearly cognitively driven (that is, it is shaped by knowledge, beliefs, expectations, needs etc.), it operates before the onset of perceptual processing, and therefore, does not imply the cognitive penetrability of perception. One could say at most that cognition indirectly affects perception, in the sense that the modifications in perceptual circuitry are connected to cognitive factors mediated by attention. This is an indirect form of cognitive penetrability of perception, in that the contents of our cognitive stances do not affect the kind of the neural modifications but only determine, as it were, the conditions of learning by means of attentional mechanisms. As Pylyshyn (1999) remarks, to argue that this is a form of cognitive penetrability is like arguing that, because the decision to wear glasses is cognitively determined and because wearing glasses affects perception, perception is cognitively penetrable. We will discuss in the next section the philosophical implications of the distinction between direct and indirect cognitive penetrability.

So, the perceptual systems are to some extent plastic, as Churchland argues. But this plasticity is not the result of the penetration of the perceptual modules by higher cognitive states, but rather, the result of learning-induced changes that are modulated by the retinal input and task-demands. Fodor (1983), given his view that the perceptual modules have a fixed architecture, had to concede that if evidence is found for diachronic changes in the functional architectures of the modules, then the modularity of perception would collapse. But this is not necessarily so.

First the data-driven changes can be accommodated by the notion that the modules are semi-hardwired. All this view requires is that the functional changes reshape the microcircuitry and not the macrocircuitry of the modules. Bearing in mind that priming enhances performance, one cannot see how such learning could reshape their basic macrocircuitry. Second, even though the perceptual systems do not have a fixed architecture, the factor that modulates the rewiring is task-driven and not cognitively driven. This bars any movement from the possibility of rewiring of the perceptual systems to the cognitive penetrability of these systems, and thus, to the incommensurability of scientific theories.

Philosophical Implications

Let me now turn to the implications of the possibility of learning-induced changes in the visual system as these relate to the issue of the existence or not of a theory-neutral basis on which the issue of rational choice among scientific theories and scientific relativism rest. The question boils down to whether scientists with different experiences could form a different *perception* of the same retinal image. Suppose that, as a result of learning through repeated experience in her field, a scientist has somewhat shaped her perceptual sensitivity according to her specific professional needs and can recognize patterns that others cannot. She has learned which dimensions of visual analysis to attend to, and this process has reshaped her basic sensors by selecting the output of certain feature detectors. Suppose further that this learning has induced changes in the circuitry of her early vision, altering her visual perception (the part of vision, which is supposedly cognitively impenetrable). Hence, the answer is *yes*; some scientists who are trained in certain fragmented patterns and have stored them in the so-called "priming memory" may be able to recognize patterns that others could not. Suppose further that these changes affect her assessment of experiential evidence about theory evaluation.

Does this pose a threat to the possibility of creating a theory-neutral perceptual basis, and thus, does it constitute a basis on which the incommensurability of scientific theories could be established? I think that it does not, since as I have argued, this neural change is task or data-driven and not theory-driven. The difference is an important one for the following reasons:

First, all humans have roughly similar perceptual circuits (barring some damage or other). Thus, despite the fact that no two humans share identical brain circuits, we all cut the world in roughly similar ways. We all share, for instance, the same neural mechanisms for perceiving colors, and thus, we have the same conceptual representations of colors (Barsalou, 1999; Lakoff, 1987). Rosch's (Rosch, et. al., 1976) findings that there exists a "universal" basic-level categorization of objects in the world, which in the case of living things corresponds to the categorization into natural kinds, seem to confirm the contention that humans cut, at some level of analysis, the world roughly in the same way. The existence of universal natural kinds can be attributed to many causes, one of which is that the animals that belong to the same natural kind have roughly the same overall shape. Since shape is one of the attributes that matters the most with regard to the human-environment interaction, shape plays an important role categorization. The fact that we all cut the world into the same natural kinds supports the thesis that we all perceive shapes the same way, undoubtedly because we share (ceteris paribus) the same visual circuits.

Second, all scientists have had experiences of more or less the same objects; they share more or less the same scientific education, and work with roughly the same objects and instruments. Thus, their brains share a roughly similar basic microcircuitry, as far as this circuitry bears on the practice of their profession, since the circuitry is formed as a result of experience.

Third, even if some of them have acquired some particular priming memory, and as a result can *perceive* patterns that others cannot, nothing precludes the latter from undergoing the same training and reestablishing a common *perceptual* ground. Learning of this kind is data- and task-driven, which means that the same training will almost certainly produce the same "priming memory". It is at this point that the difference between the direct cognitive penetrability of perception by beliefs and expectations and the indirect penetrability through task-driven learning, which in its turn is shaped by cognitive is cashed out. In task-driven learning, cognition indirectly mediates the process through the allocation of attention. Attention, can be controlled though, since people can be instructed to focus their attention on such and such a location and scan for such and such a feature, despite the fact that these people may have entirely different intentional stances. Once this factor has been controlled, differences in beliefs etc., do not affect the course of the "priming" training, and thus, of perceptual learning. This implies that similar training will induce similar brain changes. Thus, experience-induced plasticity of the brain does not threaten the possibility of a theory-neutral perceptual basis.

The difference between data-driven and theory-driven learning in general is important. The task and what one should attend to can be specified

intersubjectively in groups with varying theoretical commitments. Since the whole enterprise is data and task-driven, the same task is bound to induce similar changes. This allows scientists to perceive the same things after some training, even if initially one of them was more capable than the other to perceive certain patterns. This way a channel of communication is established, since now they perceive similar things, no matter how they interpret them, which importance they attribute them and so forth. This explains why scientists working within very different paradigms can test one the experiments of the other, compare their results etc., even though they may disagree as to their importance and confirmatory role, a finding that receives ample support from research in the history of science (Gooding, 1990; Nersessian, 1984). This finding shows that even though different theoretical frameworks shape the design of experiments and their interpretations still scientists within different paradigms can understand what others scientists are doing.

By introducing a distinction between a bottom-up and non-semantic perception and a semantic cognition I join a long tradition of similar distinctions. Jackendoff (1989) distinguishes "visual awareness" from "visual understanding". Similarly Dretske (1995) distinguishes a "phenomenal sense of see" from a "doxastic sense of see". To the extent that the first parts of the pairs clearly correspond to a non-epistemic sense of perception, and the second parts of the pairs to an epistemic sense of perception, these distinctions are coextensive with the "perception-observation" distinction that I introduced in the introduction.

I would like to close the discussion regarding the philosophical implications of perceptual learning with a remark on the dichotomy between perception and cognition. In the introduction I defined perception as the set of processes that transform sensations to cognitively usable structures, and distinguished between perception and cognition, by claiming that the former is bottom-up, whereas the latter is not. This dichotomy however, should not be taken to imply a functional and even a neural distinction between perception and cognition.

I have argued elsewhere (Raftopoulos, 2001a; 2001b) that many cognitive functions (e.g., imagery and spatial conceptualization) take place at the same neural areas that support early vision (see also Barsalou, 1999). In this sense, the mechanisms implementing perception and cognition cannot be divorced. Since the perceptual input systems are necessarily involved in higher cognitive tasks, our conceptual systems are severely constrained by the architecture of the perceptual modules. Perception does not serve only as the faculty that provides input to higher cognition and then comes on-line, after the cessation of the conceptual processing, in order to test empirically its outcome, but also constitutes an active participant of the conceptual processing itself.

This does not, mean, however, that perception and cognition function simultaneously, as Barsalou (1999) claims. There is ample neurospychological and neurophysical evidence suggesting that the perceptual processes precede cognitive processes of a scene, and that their outcomes differ. Thus, it makes sense to distinguish between perceptual and cognitive processes, even though cognition should be extended to include perception. But even if perceptual systems are cognitively penetrable, still a case cannot be made for incommensurability. For, according to Barsalou, the top-down information is overridden if in conflict with the bottom-up information coming from the perceptual modules. Thus, given some incoming information, different cognitive stances cannot cause different perceptions of a visual array.

Conclusion

I have argued in this paper that perceptual learning and the resulting rewiring of the early visual systems need not suggest the cognitive penetrability of early vision, since this form of learning is experience and task-driven and not theory-driven. This is so because perceptual learning, by being modulated by attention, is only indirectly affected by cognition. To the extent that attention can be controlled, the influence of cognition on early vision is neutralized.

We see the problem in the arguments against the cognitive encapsulation of perception. In attempting to demonstrate the cognitive penetrability of our perception and that the theoretical neutrality of observation is false, they confuse the plasticity of the brain and perceptual learning with cognitive penetrability. But the former does not entail the latter. The only way out is to argue that the experience-induced learning changes the way we observe the world, and this, in its turn, by means of some top-down flow of information which affects the way we perceive. Though it is true that our experiences shape our theories and the way we see the world, to say that these theories influence the way we *perceive* the world is question begging, since one must show that this top-down influences occur.

References

Ahhisar, M., and Horchstein, S (1993). Attentional control of early perceptual learning. *Proceedings of the National Academy of Science*, (pp. 5718-5722),

USA, 90.
Antonini, A., Strycker, M. P., and Chapman, B. (1995). Development and plasticity of cortical columns and their thalamic inputs. In B. Julesz and I. Kovacs (Eds.) *Maturational windows and adult cortical plasticity*. Reading, MA: Addison-Wesley.
Barsalou, L. (1999). Perceptual symbol systems. *Brain and Behavioral Sciences*, *22*, 577-660.
Churchland, P. M. (1988). Perceptual plasticity and theoretical neutrality: A reply to Jerry Fodor. *Philosophy of Science*, 55, 167-187.
Churchland, P. (1989). The anti-realist epistemology of Van-Fraassen's *The Scientific Image*. In B.A. Brody and R. E. Grandy (Eds*.) Readings in the philosophy of science*, Englewood Cliffs, N.J: Prentice Hall.
Dretske, F. (1995). *Naturalizing the mind*. Cambridge, MA: The MIT press.
Egeth, H. E., Virzi, R. A., and Garbart, H. (1984). Searching for conjuctively defined targets", *Journal of Experimental Psychology*, *10*, 32-39.
Fodor, J. (1983). *The modularity of mind*. Cambridge, Mass: The MIT Press.
Goodale, M. A. (1995). The cortical organization of visual perception and visuomotor control. In S. M. Kosslyn and D. N. Osherson (Eds.), *Visual cog*nition (Vol. 2). Cambridge, MA: The MIT Press.
Gooding, D. (1990). *Experiment and the making of meaning*, Kluwer Academic, Dordrecht.
Gregory, R. (1974). *Concepts and mechanisms of perception*. New York: Charles Scribners and Sons.
Greenough, W. T., Black, J. E., and Wallace, C. S. (1993). Experience and brain development. In M. H. Johnson (Ed.), *Brain development and cognition: a reader*. Cambridge: Basil Blackwell.
Hanson, N. R. (1958). *Patterns of Discovery*. Cambridge: Cambridge University Press.
Hildreth, E. C., and Ulmann S. (1989). The Computational study of vision. In M. I. Posner (Ed.), *Foundations of Cognitive Science*. Cambridge, MA: The MIT Press.
Hoffman, J. E. and Subaramaniam, B. (1995). Saccadic eye movements and selective visual attention. *Perception and Psychophysics*, *57*, 787-795.
Jackendoff, R (1989). *Consciousness and the computational mind*. Cambridge, MA: The MIT Press.
Kahneman, D., and Treisman, A. (1992). The reviweing of object files: Object-specific integration of information", *Cogntive Psychology*, *24*(*2*), 175-219.
Karni, A., and Sagi, D. (1995). A memory system in the adult visual cortex. In B. Julesz and I. Kovacs (Eds.) *Maturational Windows and Adult Cortical Plasticity*. Reading, MA: Addison-Wesley.
Kawabata, N. (1986). Attention and depth perception. *Perception*, *15*, 563-572.
Kuhn, T. S. (1962), *The structure of scientific revolutions*. Chicago: Chicago University Press.
Lakoff, G. (1987*). Women, fire, and other dangerous things*. Chicago; Chicago University Press.
Marr, D. (1982), *Vision: A Computational investigation into human representation and processing of visual information*. San Francisco, CA: Freeman.
McLeod, P., Driver, J., Dienes, Z., and Crisp, J. (1991). Filtering by movement in visual search", *Journal of Experimental Psychology: 17*, 55-64.
Meissirel, C., Dehay, C., and Kennedy, H. (1993). Transient cortical pathways in the pyramidal tract of the neonatal ferret. *Journal of Comparative Neurology*, *338*, 193-213,
Nersessian, N. J. (1984). *Faraday to Einstein: Constructing Meaning in Scientific Theories*. Hingham, MA: Martinus Nijhoff Publishers.
Peterson, M. A., and Gibson, B. S. (1991). Directing spatial attention within an object: Altering the functional equivalence of shape descriptions. *Journal of Experimental Psychology*, *17*, 170-182.
Pylyshyn, Z. (1999). Is Vision Continuous with Cognition? *Behavioral and Brain Sciences*, *22*, 341-365.
Raftopoulos, A. (2001a). Is Perception Informationally Encapsulated? The Issue of the Theory-Ladenness of Perception. Forthcoming in *Cognitive Science*.
Raftopoulos, A. (2001b). Reentrant Pathways and the Theory-Ladeness of Observation. Forthcoming in *Philosophy of Science*.
Rosch, E., Mervis, C., Wayne, G., Johnson, D., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382-439.
Stiles, J. (1995). Plasticity and development. Evidence from children with early occurring focal brain injury. In B. Julesz and I. Kovacs (Eds.) *Maturational windows and adult cortical plasticity*. Reading, MA: Addison-Wesley.
Ulmann, S. (1979). *The Interpretation of visual motion.* Cambridge, MA: The MIT Press.
Yeo, R. M., Yonebayashi, Y. , and Allman, J. M. (1995). Perceptual memory of cognitively defined contours: A rapid, robust and long-lasting form of memory. In B. Julesz and I. Kovacs (Eds.) *Maturational windows and adult cortical plasticity*. Reading, MA: Addison-Wesley.

The influence of semantics on past-tense inflection

Michael Ramscar (michael@cogsci.ed.ac.uk)
University of Edinburgh, 2 Buccleuch Place
Edinburgh, EH8 9LW, Scotland.

Abstract

Previous theories of past-tense verb inflection have considered phonological and grammatical information to be the only relevant factors in the inflection process (e.g. Bybee & Moder, 1983; Rumelhart & McClelland, 1986; Kim, Pinker, Prince & Prasada, 1991). This paper presents three experiments that show that semantic information plays a decisive role in determining the inflection of both existing and novel homophone verb stems. These findings indicate that regular and irregular inflections are determined by semantic and phonological similarities in memory, and furthermore that people are not responsive to the kind of grammatical distinctions amongst verb roots that default rule theories of inflection (Pinker, 1999) presuppose.

Introduction

In most theories -- and studies -- of past-tense verb inflection, phonological and grammatical information have been considered to be the two relevant factors in the inflection process (e.g. Bybee & Moder, 1983; Rumelhart & McClelland, 1986; Kim, Pinker, Prince & Prasada, 1991; Pinker, 1991; 1999). However, in some models of inflectional processing (MacWhinney & Leinbach, 1991; Joanisse & Seidenberg, 1999), semantic processes have been included to help explain the processing of homophone verbs (e.g. *brake*/*break*). Since *brake* and *break* both sound the same, phonology alone cannot distinguish which of *broke* or *braked* is to be the correct past tense form for the input breɪk.

Although using semantic information to guide this process appears intuitively plausible, it has not been supported empirically, and indeed this suggestion has been fiercely criticised by Pinker and colleagues (Kim et al, 1991; Pinker, 1999), who put forward an alternative, natavist account of homophone inflection (Pinker, 1991; 1999). This predicts that the regularisation of irregular sounding verb stems is driven by innate grammatical sensitivity: verbs that are instinctively perceived to be denominal will be automatically regularised. This account is supported by results reported by Kim et al (1991) which indicate that grammatical factors correlate better than semantic factors with people's ratings of the acceptability of past tense forms in context, although these results do not rule out any semantic role in inflection.

So do semantics have any influence on the past tense forms speakers produce? This paper seeks to clarify and directly address this question.

Semantics and past-tense inflection

To initially test whether semantic similarity can affect the inflection of verb past tenses, Experiment 1 examines the past tense forms native English speakers produce for novel (nonce) English verbs whilst holding phonological factors constant and varying the semantic contexts in which the verbs are presented. The phonologically similar nonces *sprink* and *frink* are presented in contexts that primed either the existing phonologically similar regular forms *blink* or *wink*, or the existing phonologically similar irregular form *drink*. It is hypothesized that if semantic similarity played a part in the inflection process, then there will be significant differences between the proportion of regular and irregular forms produced, in line with whether the semantic context favored an existing regular or irregular verb.

Experiment 1

Participants. Participants were 120 visitors a shopping mall in Edinburgh, Scotland, and 40 students at the University of Edinburgh. All were native English speakers and participated voluntarily in the study

Materials. A set of cards were printed with a paragraph that contained a highlighted nonce verb (*sprink* or *frink*) in a context in which the nonce was in its infinitive tense, and a blank that later required its past tense. Two of the contexts were further manipulated so that they primed either two existing regular verbs -- *blink* and *wink* -- that are phonologically similar to the nonces, or an existing irregular verb -- *drink* -- that is phonologically similar to the nonces. The contexts constructed are shown in Table 1: in the *drink* context, the nonce was shown in a context that used it as a verb to describe the consumption of vodka and fish, whereas in the *blink* and *wink* context the nonce was shown in a context that used it as a verb to describe a symptomatic affliction of the eye-lid associated with a fictitious disease.

A third context was designed to semantically prime neither *drink* nor *blink* or *wink* (instead the nonce was used as a verb describing a hypnotic trance and was semantically similar to the regular verb *meditate*; see Table 1), whilst a control presented the nonce in a context that provided few semantic clues ("John likes to *frink*. Last week he _____").

In order to control the phonological properties of the nonces in the semantic contexts, both the initial presentation of the nonce, and the blank which was used to elicit the past tense form from participants were embedded in the same sentence substructure in each of the three

semantic contexts. Each nonce and blank was preceded by at least two identical words (containing at least three identical syllables), and succeeded by at least one identical word (containing at least one syllable).

Table 1. Examples of the semantic context passages used in Experiment 1. The nonce (in this case, *frink*) is italicized. The text highlighted in bold was used to phonologically control the presentation of the nonce and then later the blank, and is identical in all of the contexts.

In a traditional spring rite at Moscow University Hospital, the terminally ill **patients all** *frink* **in** the onset of good weather, consuming vast quantities of vodka and pickled fish. In 1996, his favourite vodka glass in hand, cancer **patient Ivan Borovich** _____ **around** 35 vodka shots and 50 pickled sprats; it is not recorded whether this helped in his treatment.

Passage 1 - irregular context - primes *drink*.

In a classical symptom of Howson's syndrome, **patients all** *frink* **in** their right eye if they are left handed or left eye if right handed, their eyelids opening and closing rapidly and uncontrollably. In 1996, in extreme discomfort due to his bad eye, Howson's **patient Ivan Borovich** _____ **around** 35 times per minute for two days, causing severe damage to the muscles in his left eyelid.

Passage 2 - regular context - primes *blink* and *wink*

In a controversial alternative therapy at Moscow University Hospital, the terminally ill **patients all** *frink* **in** the afternoons on alternate days, going into a trance-like state that lowers the heartbeat to alleviate pain. In 1996, emitting a steady, low humming sound, cancer **patient Ivan Borovich** _____ **around** two weeks or so (the nurses lost count!) without a day off. Afterwards, doctors claim, his cancer was cured.

Passage 3 - context primes neither *drink* nor *blink* and *wink* but rather is similar to *meditate*

Procedure. Participants were verbally briefed to "read a piece of text. As you read through the text, you will see a novel word that has been highlighted, and later a blank, where a word has been left out. We would like you to tell us the form of the highlighted new word that you think is appropriate to the context in which you find the blank. It is important that the form you choose matches the context of the sentence." Participants were also told to "concentrate on how the new form of the novel word 'sounds' in the context, not on how it might be spelled". After this briefing, participants were given a card containing the example "A single *wucterium* can be very dangerous. When they breed and multiply, a build up of _____ can prove lethal" and verbally informed that they might choose to fill in the blank with, *wucteriums* or *wucteria*, or anything else, depending on which seemed appropriate to them. These cards were then retrieved by the experimenter, and removed from the sight of the participant. Once this was done, the participant was given either a *frink* or a *sprink* card. Participants read the passage on the card and produced a verbal response to the fill-in-the-blank inflection task.

Participants verbalized all their responses, which were transcribed by the experimenter.

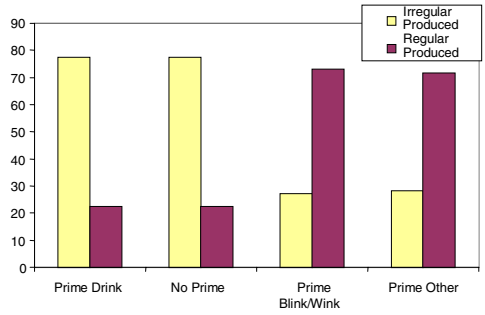


Figure 1. Percentages of each form of inflection for each condition in Experiment 1.

Results. The results obtained were consistent with the hypothesis that semantic similarity could affect the inflection of the past tenses of nonce English verbs when phonological similarity constraints were satisfied. Overall, 96.3% of participants produced a recognizable past tense form (3.7% produced inflected forms of the active past perfect progressive form, adding +/ *ing* to the nonce stem, and were discarded). Of those participants that produced past tense forms, 75.3% responded in a prime-consistent manner (see figure 1). The bias towards producing a form consistent with the past tense form of the existing verb primed by the semantic context was significant in a 2x2 chi-square analysis, $\chi^2(1, N=77) = 19.669$, $p<.0001$). Analyses of the individual nonces also showed the consistency bias to be significant: *sprink*, $\chi^2(1, N=39) = 9.776$, $p<.005$); *frink*, $\chi^2(1, N=38) = 7.204$, $p<.01$).[1]

Comparing the responses of participants who had seen the nonces in the context that primed *drink* to those who had inflected the nonces in the semantic context that primed neither *drink* nor *wink/blink* -- but rather indicated that the nonce verbs described a meditative state -- showed a significant trend towards regularization; 71.8% of participants in this condition produced a regular past tense form.

On the other hand, in the control condition in which participants encountered the nonce in a semantically neutral setting, 77.5% of participants produced an irregular past-tense. Comparing this to the 73% of participants produced *regular* form for the past tenses of *frink and* sprink when they encountered them in a context that primed the regulars *blink* and *wink* indicated a tendency to regularize in the latter case: $\chi^2(1, N=77) = 15.901$, $p<.0001$; *frink*, $\chi^2(1, N=38) = 5.743$, $p<.05$; *sprink*, $\chi^2(1, N=39) = 7.621$, $p<.01$). Similarly, comparing the neutral respondents to participants who encountered the nonces in a context that did not prime drink, but which did prime existing regulars such as *meditate* and *heal*, there was also a significant increase in the number of regular forms produced $\chi^2(1, N=79) = 15.5$, $p<.0001$; *frink*, $\chi^2(1, N=38) = 7.204$, $p<.01$; *sprink*, $\chi^2(1, N=40) = 7.621$, $p<.01$), see figure 1. This suggests that the

[1] The values given for the analysis of the individual nonces use Yates' corrected chi-square.

significant effect produced by semantically priming *wink* and *blink* was to increase the proportion of regulars produced, and that this effect was maintained when other regular forms were contextually primed.

Discussion. This experiment set out to examine whether semantic similarity might play a complimentary role to phonological similarity in inflection, i.e. whether it might have an influence on the form of semantically similar words alongside phonological constraints. The results obtained suggest that when people encounter a novel verb form that is phonologically and semantically close to an existing verb form, then the likelihood is that they will use the same pattern of inflection to form the past tense of the novel verb form as is used to inflect the past tense of the existing verb form. This finding neatly compliments the findings of Bybee and Moder (1983), who discovered that encountering a novel verb form that is phonologically similar to a cluster of phonologically similar irregulars increases the likelihood that the nonce verb form will be inflected irregularly. Once phonological similarity constraints have been met, semantic information also appears to play a role in inflection.

Semantics and grammatical analysis

The results of Experiment 1 suggest that semantic information does influence the past tense inflection process: the past tense forms of inputs that are phonologically similar to phonologically similar regular and irregular verb forms can be significantly influenced by semantic information. Given that this finding suggests that semantic information is at least sufficient for resolving the inflection patterns of homophone verbs, it raises the question of whether semantic information is also necessary for this to occur. Are semantics always used to determine the inflection of homophone verb forms? Or is other information, such as the grammatical status of the verb inputs, as suggested by Pinker and colleagues (Kim et al, 1991; Pinker, 1999) also sufficient to determine the outcome of this process?

To examine these questions, Experiment 2 was designed to test the hypothesis that semantic similarity, and not formal grammatical analysis, would be the important determining factor in the inflection of homophone verbs. According to Kim et al (1991; see also Pinker, 1999), in carrying out inflection people perform formal grammatical analyses on lexical input. As a result of these analyses, only verbs with verb roots (deverbal verbs) will be given an irregular past tense form; verbs with noun roots (denominals) will be inflected regularly.

> "People should regularize headless [denominal] forms only when they *perceive* the words to be headless… they should have a sense of when a word is based on another word... When they don't – when they are oblivious to the noun in a verb-from-a-noun and imagine that it is just a stretched verb root – the theory predicts that they *should* stick with the irregular." (Pinker 1999, p. 171.)

The findings obtained so far support the view that inflection is carried out as a result of competing constraints (involving phonology, semantics and word frequency) in memory. If such a similarity-based process was determining the inflection of verbs, it was predicted that when nonce verbs were presented to participants as denominal verbs in a context that semantically primed an irregular verb cluster, then they would be inflected irregularly (contrary to the 'formal grammatical analysis' hypothesis).

Experiment 2

Participants. Participants were 80 visitors to a shopping mall in Edinburgh, Scotland and 104 students at the University of Edinburgh. All were native English speakers and participated voluntarily in the study.

Materials

Inflection Task. The materials used in this study were modified versions of the cards used in experiment 2. For the purposes of this experiment, the contexts were modified so that the first sentence in each introduced the nonce as a noun, so that subsequent use of the nonce was clearly as a denominal verb. Kim et al (1991, p. 207) claim that this information is a "necessary and sufficient condition for the regularization effect." In the *drink* context, the nonce was shown in a context that used it as a noun (a kind of tapa comprising vodka and fish) and later as a verb to describe the consumption of the tapa (and by extension, vodka and fish), whereas in the *blink* and *wink* context the nonce was shown in a context that used it as a noun to describe a muscle, and later as a verb to describe a symptomatic affliction of that muscle. In the *meditate* context, which was designed to semantically prime neither *drink* nor *blink* or *wink* (the nonce was used as a verb describing a hypnotic trance), the nonce was first presented as the name of a Siberian religious sect that mediated.

Grammatical Analysis Task. 60 participants completed a questionnaire as part of a grammatical analysis measurement condition. Participants were given booklets containing one example of a "deverbal" nonce target context, one example of a "denominal" nonce target context and six other contexts relating to existing verbs. The verbs in these contexts were loosely pre-classified as either deverbal, denominal, or obscure but discenably denominal. The contexts were designed to provide participants with sufficient information to facilitate their making a meaningful denominal/deverbal analysis.

Procedure

Inflection Task. The procedure for this task was the same as for Experiment 1, above.

Grammatical Analysis Task. The presentation order of the contexts was randomized and ordered so that each participant saw one nonce denominal context and one nonce deverbal context from different scenarios, along with six other items. Participants were asked to indicate whether in their judgement, the verb was 'being used in a normal

'verb-like' way, or whether it is being used as a verb in relation to a noun. To take a word like 'drink' for example, in (1) 'John likes to drink beer'... you might decide that "drink" is being used as an ordinary verb. But, in the example… (2) It's always a good idea to relax your guests. Whenever guests arrive at my house, I immediately snack them and *drink* them. I find that refreshments set them at ease.' You might decide that 'drink' is being derived from the noun 'drink' (such that it means something like 'to serve drinks.'" Participants gave their ratings on a 7 point scale where 1 = definitely normal 'verb-like' use and 7 = definitely 'verb from noun' use.

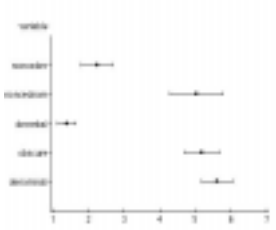


Figure 2. Means plot for the verb categories in the grammatical analysis condition in Experiment 2. The upper 2 plots show mean ratings for the nonces in denominal (noncedenom) and deverbal (noncedev) contexts. The bottom 3 plots show means rating for the existing verb categories

Results. The results obtained were consistent with the hypothesis that semantic similarity was affecting the inflection of the past tenses of nonce English denominal verbs when phonological similarity constraints were satisfied; they were not consistent with any effect of formal grammatical analysis affecting inflection.

Of those participants who encountered a novel denominal verb in a context that primed an existing irregular verb (i.e. in a context that made it appear semantically similar to that verb), 72.5% produced an irregular past tense for *sprink* or *frink*, whilst of those participants who encountered the novel denominal verb in a context that primed an existing regular, 75% produced a regular past tense for *sprink* or *frink* (this prime consistency effect was significant: $\chi^2(1, N=80) = 18.061$, $p<.0001$). The context which described a meditative or hypnotic state) also promoted a regular inflection of *sprink* and *frink* (71.8%).

The results of the grammatical analysis task are summarized in figure 2. It appears that that participants could distinguish the grammatical origins of the various verbs in these contexts (this is reflected by the variance in category means for the 'regular', nonce and obscure denominals with respect to the regular and nonce deverbals). A one way ANOVA indicated the variance in the scores assigned by participants to the verb categories with respect to their grammatical origins of the categories was significant $F(4) = 74.65$, $p<0.0001$. T-tests between the mean scores for the denominal and deverbal nonces in context showed that participants had detected significant differences in their grammatical origins $t(56) = 6.45$, $p<.0001$ (individual nonces in contexts: *nonce - drink*, $t(17) = 3.47$, $p<.005$; *nonce - wink/blink*, $t(17) = 3.19$, $p<.01$; *nonce - meditate*, $t(18) = 3.26$, $p<.005$). These findings indicate that the inflection results do not stem from any failure on the part of participants to perceive the grammatical categories of the nonce verbs.

Discussion. This experiment set out to establish whether semantic similarity or formal grammatical analysis would influence the inflection of novel denominals. The finding that patterns of inflection were consistent with semantic priming -- as found in experiment 2 -- is entirely consistent with the hypothesis that semantic similarity is an important constraint in determining inflection when phonological constraints have been met. Moreover, the fact that a majority of irregular forms were produced for both denominal nonces when the context they were presented in suggested that they were semantically similar to an existing irregular verb directly contradicts the formal grammatical analysis hypothesis, which predicted that because these verbs had a noun root they would be inflected with a regular past tense form.

Semantics or Grammar?

The evidence accrued against grammatical analysis in Experiment 2 appears to conflict quite drastically with the findings of Kim, Pinker, Prince & Prasada (1991) who report that participants in their experiments did use grammatical analysis in determining the past tenses of verbs. As noted above, an experiment by Kim et al (1991) appeared to show that grammatical analysis is a better predictor of past tense forms than semantics. Since these findings are clearly incompatible with the findings of Experiment 2, it seems worth considering in some detail this earlier study.

The most important thing to note about the experiment by Kim et al (1991) that sought to discriminate between the two competing accounts of homophonic inflection -- grammatical analysis versus semantics -- is that judgments of the grammatical status of verbs (whether they were denominal or deverbal) were not collected from impartial observers. Instead, the experimenters relied on their own intuitions for deciding which verbs are denominal and which deverbal. This practice raises several concerns when Kim et al's materials are subject to close scrutiny. To take one example, Kim et al classify the verb *to lie* (as in confabulate) as denominal (in their materials, "Sam always tells lies when he wants people to think he's better than he really is, He *lied/lay* to me last night about how good a golfer he is"). Presumably, this is because one can tell *a lie*; a *liar* is someone who tells *lies*, etc. However, Kim et al simultaneously classify the verb *to drink* as deverbal (presumably, given that *drink* is an irregular, and the grammatical analysis hypothesis suggests that all denominals will be regular, Kim et al assumed *drink* had to be deverbal). Yet it would appear that *any* reason that one can think of for suggesting that *lie* is denominal applies

equally to *drink*: one can drink a *drink*; a *drinker* is someone who imbibes *drinks* etc. Yet these factors -- and the potentially flawed assumptions behind them -- are critical to Kim et al's analysis of their results. If participants were to judge the past tense of *lie* as *lied* and the past tense of *drink* as *drank* (as indeed they did in Kim et al's experiments), then this would count as evidence *for* the grammatical analysis hypothesis in Kim et al's subsequent analysis. Grammatical analysis would be credited with predicting that *drink* is irregular whereas *lie* must be regular, even though as far as one can see, the only reason that *lie* was judged by Kim et al to be a denominal and *drink* a deverbal was because the former was a regular and a latter an irregular in the first place.

In the light of these considerations, the following experiment was designed to re-examine the findings of Kim et al (1991), subjecting both putative mechanisms for predicting the inflection of homophonous verbs to the same standard of testing. Participants in a semantic reminding condition were presented with a target verb in context, and an example of a homophonous irregular verb (the examples had been adjudged by another group of participants in a pre-trial to be very typical uses of the irregular). Participants were asked to rate the extent to which the action or activity associated with the verb in the target context reminded them of the kind of activity or action picked out by the example verb; participants were encouraged to think of all of the activities they usually associated with that use of the target. The semantic hypothesis tested in Experiment 3 was simply that these semantic reminding judgements would be a good predictor of the acceptability of irregular past tense forms.

In contrast to this, the grammatical analysis hypothesis tested in this experiment predicts that verbs that subjects' perceive to be denominal will be regularized, thus the grammatical analysis hypothesis would expect the grammatical analysis of verbs as being denominal would be a good predictor of the acceptability of regular past tense forms (see Kim at al, 1991).

Experiment 3

Participants. Participants were 101 native English speaking undergraduate students at the University of Edinburgh who participated voluntarily in the study.

Materials. 24 contexts were used to present 12 phonologically similar verb pairs in the present tense, (e.g. "Charlie Wilson of United is a real prima donna. He never gets on with the game. Instead, he just shows off. He tries to *grandstand* all the time, and it really gets on my nerves."). In each context, the target verb was italicized and underlined. A second set of 24 contexts presented the target word first in its present tense, and later as either a regular or an irregular past tense (e.g. " Charlie Wilson of United is a real prima donna. He never gets on with the game. Instead, he just shows off. He tries to grandstand all the time, and it really gets on my nerves. In the game with Rovers on Saturday he got an early goal and *grandstood* [or *grandstanded*] the rest of the match.").

Procedure. 3 groups of participants were used to obtain three rating measures: semantic reminding, grammatical analysis and acceptability of past-tense form.

Semantic Reminding. Participants were presented with an example 'typical use' of the irregular form of the homophonic verb for comparison purposes (e.g. " The soldiers were told to *stand* at ease."). Each of the examples had been rated by 12 participants in a pre-test for typicality and achieved a mean score of > 5.8 on a scale where 1 = not at all typical and 7 = very typical. Participants were instructed to compare the comparison verb (e.g. stand) and the highlighted verb in the target (e.g. *grandstand* above) and to rate the extent to which "the activity or action described by the *underlined* word … -- *taken in the whole context* -- remind[ed… them] of the comparison word." Participants were additionally instructed to " to consider all the possible things [they] *usually* associate with this use of the word." Ratings were given on a 7 point scale where 1 = strong reminding and 7 = no reminding.

Grammatical Analysis. Participants were asked to indicate whether a verb was 'being used in a normal 'verb-like' way, or whether it is being used as a verb in relation to a noun.' Participants gave their ratings on a 7 point scale where 1 = definitely normal 'verb-like' use and 7 = definitely 'verb from noun' use.

Acceptability Of Past-Tense Form. The instructions in this task mirrored those in Kim et al (1991). Participants were told to "to concentrate on how the words 'sound' in their context, as you read them, not on how they might be spelled." Participants were asked to indicate "how likely it is that the form you have seen is the correct one for that context. By 'correct', we mean the one that other native English speakers would naturally and intuitively use (i.e. the form - or sound - that comes most naturally to you)." Participants were also asked to note that correct did not "refer to the kind of 'proper' English that gets taught in grammar lessons or style manuals." Participants gave their ratings on a 7 point scale: 1=not acceptable; 7=highly acceptable.

Results. A multiple regression analysis of the relationship between semantic reminding, grammatical analysis and irregular past tense acceptability indicated that both predictor variables accounted for 68% of the total variance in the irregular rating scores ($F(1, 21)=25.14$, $p<.0001$). For comparison purposes, a regression analysis of semantic reminding to irregular past tense acceptability was performed, which accounted for 67% of the total variance in the irregular rating scores: $F(1, 22)=47.71$, $P<.0001$), indicating that grammatical analysis uniquely accounted for only 1% of the total variance observed. A regression analysis of grammatical analysis to irregular past tense acceptability accounted for 37.5% of the total variance in the irregular rating scores: $F(1, 22)=14.87$, $p<.001$),

indicating that semantic reminding uniquely accounted for a significant 29.5% of the total variance observed.

A multiple regression analysis of the relationship between semantic reminding, grammatical analysis and regular past tense acceptability indicated that both predictor variables accounted for 36% of the total variance in the regular rating scores ($F(1, 21)=7.479$, $p<.005$). Again as a comparison, a regression of semantic reminding to regular past tense acceptability was performed, which indicated that semantics alone could account for 35% of the total variance in the regular rating scores: $F(1, 22)=13.52$, $p<.005$), indicating that again, just 1% of the observed variance could be uniquely accounted for by grammatical analysis. A regression analysis of grammatical analysis to regular past tense acceptability accounted for 1.7% of the total variance in the irregular rating scores: $F(1, 22)=1.397$, $p>.2$), indicating that semantic reminding uniquely accounted for a significant 33.3% of the total variance observed

Discussion. In the light of some potential flaws identified in the study comparing the predictive power of semantics versus grammatical analysis reported by Kim et al (1991), this experiment set out to re-examine those previous findings using a suitably modified experimental paradigm. Strikingly, once both predictor variables were subject to the same standard of empirical scrutiny, the predictive effect of grammatical analysis on past-tense acceptability reported by Kim et al in their study has almost entirely vanished. In this study unconfounded grammatical analysis predicted only a very marginal, insignificant amount of the data collected in the past-tense rating condition. On the other hand, it appears that semantic reminding is a good predictor of the acceptability of inflected forms. Semantic reminding uniquely predicts a significant proportion of participants' irregular and regular past tense acceptability scores (29.5% and 33.4% respectively).

General Discussion

Inflectional morphology has become the example domain -- and hence the battleground -- for wider debates about the nature of linguistic knowledge and knowledge representation, and in particular connectionist versus symbolic (and particularly, rule-based) theories of mental representation (Pinker & Prince, 1988; MacWhinney & Leinbach, 1991; Pinker, 1999). Although single-route (connectionist) and dual-route (symbolic) accounts of inflection largely agree on the processes that determine irregular inflection (a phonological pattern associator in memory) the dual-route account differs from single-route theories by claiming that only irregular forms are processed in memory, and that a symbolic default rule determines regular inflection.

The results presented here indicate: Firstly, that semantic information derived from similarity measures in memory plays a significant role in inflection. Secondly, that these similarity measures seem to involve all verbs, both regular and irregular. Thirdly, that grammatical information regarding whether verbs are denominal or deverbal does not have a significant effect on inflection patterns. And fourthly, and finally, that semantic similarities between a nonce verb and an existing regular verb in memory can result in an increase in the amount of regular forms produced.

What is interesting about these findings is that contrary to earlier claims regarding the inability of single-route models of inflection to account for the processing of homophone verbs (Pinker & Prince, 1988; Pinker, 1991; 1999), these results are entirely compatible with a single, similarity based inflection process. On the other hand, they are not compatible with a processing model where only irregular inflections are processed according to similarity in memory. The evidence presented here regarding the role of semantics in inflection suggests that contrary to some recent claims (Pinker, 1999), a full account of the processes governing inflectional morphology has yet to be put forward. The question of whether abstract mental rules are necessary in morphological processing remains very much open; the past-tense debate is still very much alive.

Acknowledgements

I am grateful to Lera Boroditsky, Dedre Gentner, Ulrike Hahn, Marc Joanisse, Mark Steedman and Daniel Yarlett for comments on, and discussions of, this work.

References

Bybee, J. L. & Moder, C. L. (1983). Morphological classes as natural categories. *Language*, 59, 251-270.

Joanisse, M.F. & Seidenberg, M.S. (1999). Impairments in verb morphology following brain injury: a connectionist model. *Proceedings of the National Academy of Sciences*, 96(13), 7592-7597.

Kim, J. J., Pinker, S., Prince, A. & Prasada, S. (1991) Why no mere mortal has ever flown out to center field, *Cognitive Science*, 15, 173-218

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition*, 40, 121-157.

Pinker, S. (1991) Rules of language. *Science*, 253, 530-535

Pinker, S. (1999) *Words and Rules*, NY: Basic Books.

Pinker, S. & Prince, A., (1988). On language and connectionism. *Cognition*, 28, 73--193

Rumelhart D. E. and McClelland J. L (1986) On learning past tenses of English verbs. In Rumelhart D.E. & McClelland J.L (eds) *Parallel Distributed Processing*: Vol 2: Psychological and Biological Models. Cambridge, MA: MIT press.

The Emergence of Words

Terry Regier (regier@psych.uchicago.edu)
Bryce Corrigan (b-corrigan@uchicago.edu)
Rachael Cabasaan (rrcabasa@uchicago.edu)
Amanda Woodward (alw1@psych.uchicago.edu)
Department of Psychology, University of Chicago
Chicago, IL 60637 USA

Michael Gasser (gasser@indiana.edu)
Linda Smith (smith4@indiana.edu)
Cognitive Science Program, Indiana University
Bloomington, IN 47405 USA

Abstract

Children change in their word-learning abilities sometime during the second year of life. The nature of this behavioral change has been taken to suggest an underlying change in mechanism, from associative learning to a more purely symbolic form of learning. We present a simple associative computational model that accounts for these developmental shifts without any underlying change in mechanism. Thus, there may be no need to posit a qualitative mechanistic change in the word-learning of young children. More generally, words, as symbols, may emerge from associative beginnings.

Overview

Word-learning is likely to rely heavily on associative learning, such that the child comes to associate the sound "dog" with dogs, the sound "cat" with cats, and so on. However, children's word-learning abilities change significantly during the second year of life, and some have proposed that this behavioral change reflects an underlying mechanistic shift away from a purely associative base. In particular, it has been proposed that sometime during the child's second year, a conceptual insight into the symbolic, referential nature of words occurs (McShane, 1979). This insight then supports a more purely symbolic form of learning, in contrast with the simple associative learning that preceded it.

A number of changes in word-learning occur at around this age. When viewed as a totality, this array of behavioral changes does suggest a mechanistic change of some sort. We shall argue, however, that these changes may be accounted for without recourse to any posited conceptual insight, or any qualitative mechanistic change in the nature of word-learning. Instead, they flow naturally from a purely associative mechanism, operating over similarity-based representations. As these representations gradually become more peaked and finely differentiated, the child's linguistic behavior becomes more recognizably "symbolic". We argue this point by presenting an associative computational model, and demonstrating that it matches the developmental shifts of 1- to 2-year-old children. Thus words, as discrete arbitrary symbols, may emerge from fundamentally associative, similarity-based mental material.

We are not the first to propose this general idea, nor to present a computational model supporting it (Cottrell and Plunkett, 1994; Elman et al., 1996; Merriman, 1999; Plunkett et al., 1992). However, the specific cluster of behavioral issues we address have not, to our knowledge, yet been accounted for computationally.

We begin by briefly outlining the empirical evidence for a change in word-learning during the child's second year of life. We then present an associative computational model, and demonstrate that it accounts for this change. We also highlight predictions made by the model, and some preliminary evidence in favor of them. We conclude with a discussion of the ramifications of this account.

Empirical Evidence

During the second year of life, the child's word-learning behavior changes in at least four respects: ease of learning, honing of linguistic form, honing of linguistic meaning, and the learning of synonyms.

Ease of learning

As children first begin to produce words, their acquisition of new words is slow and errorful. New words are added at the rate of 1 or 2 every few weeks (Gershkoff-Stowe and Smith, 1997). Then between 18 and 22 months (when the child has about 50 words in productive vocabulary), the rate of new word acquisition accelerates dramatically, with reports from detailed diary studies of children learning as many as 36 words in a single week (Dromi, 1987). Experimental studies replicate this shift in the laboratory. At the beginning of word learning, 13- to 16-month-olds can acquire a word-object linkage in comprehension based on 4-8 training trials (Bird and Chapman, 1998). By the time children are 2 to 3 years of age, a single learning trial is sufficient for word learning in comprehension and production, and for generalization to an appropriate range of referents. Thus, children appear to shift from learning as a gradual process to the sort of all-or-none process that often characterizes symbolic learning.

Honing of linguistic form

Infants must learn what counts as a word in the language they are learning, and what does not. The developmental evidence suggests that in the beginning, words function as ordinary members of the open set of possible associates. Later, however, the range of acceptable word

forms becomes narrower. For example, Namy and Waxman (1998) found that 18-month-olds readily accepted a hand gesture as a word form — in that they learned to associate the gesture with a referent. Older children however, 26-month-olds, did not learn the association. This developmental trend has been replicated using other non-phonological "words" (Woodward and Hoyne, 1999).

There is other evidence that the process is one of "honing" or narrowing the set of possible forms. Although infants readily discriminate between individual phonemes in perceptual tasks, they do not exploit this level of detail in their initial representations of words (Stager & Werker, 1997; see also Bird & Chapman, 1998). Specifically, at the beginning of word learning, at 14 months, babies cannot learn that *bih* and *dih* refer to different items, although they can learn this for globally different forms such as *lif* and *neem*. Thus children seem to move from a state in which they are sensitive primarily to overall similarity or difference between word forms to one in which they are acutely sensitive to minor differences.

Honing of meaning

Just as forms become progressively restricted with development, so do meanings. Early in word learning, 13- and 18- month old children generalize a newly learned object name to new referents by overall similarity across all dimensions (Smith, Jones, Gershkoff-Stowe & Samuelson, 1999). But older children systematically generalize novel names for artifact-like things using the specific dimension of *shape* (Smith et al., 1999). Thus, children come to pay attention to particular dimensions of referents and disregard others - much as they do with word forms. (A distinction however is that this "shape bias" holds for object names and not other sorts of names.)

Synonyms

Children assume that two different forms carry two different meanings. This has been termed the *mutual exclusivity assumption*. One specific manifestation of this assumption is that young children tend to resist learning synonyms. Liittschwager and Markman (1994) found that 16-month-olds, who can learn a new word for an as-yet-unnamed object, have trouble learning a new word for an already-named object (i.e. a synonym). However, 24-month-olds learn novel names and synonyms equally well - they do not exhibit a particular resistance to learning synonyms. Thus, there is a shift in the ease of learning synonyms, one that occurs at about the same age as the other changes in word-learning outlined above.

These roughly simultaneous changes, in ease of learning, form-honing, meaning-honing, and synonym-learning, may suggest an underlying change of mechanism sometime near the second birthday. However, we shall argue that no qualitative change in mechanism is necessary to account for these parallel developmental trajectories. They emerge naturally from a single fundamentally associative mechanism.

Foundational Assumptions

As we have seen, children do not enter the world with a clear sense of what counts as an acceptable word form. But if this is the case, what differentiates word forms from meanings in the first place, in the mind of an infant? Both are experiences of events or objects in the world. We assume that the answer lies in the child's awareness of her interlocutor's stance as a social other. Specifically, we assume that the child will take the object of the interlocutor's *attention* as a potential referent (Baldwin et al., 1996). Further, we assume that those intentional actions of the interlocutor to which the interlocutor is *not* attending are taken as potential forms – this will include verbal utterances, gestures, and any other unattended action. It is known that pre-linguistic infants are sensitive to the object of attention of another person (Corkum and Moore, 1998). Thus, the deployment of the interlocutor's attention serves as a plausible starting-point for the development of the form/meaning distinction.

The Model

The model, shown in Figure 1, builds on these social assumptions. It accepts as input a potential form and a potential referent, which are assumed to have been determined by the interlocutor's deployment of attention. These inputs are each represented by a bank of nodes, corresponding to features of experience. Form and referent are associated in the top layer of the network, which holds a localist lexicon - in which each node stands for a distinct pairing of form and meaning. The form and meaning for a given lexicon node are encoded on its incoming weights. New nodes are added to this lexicon as new form-meaning pairings are encountered (Carpenter and Grossberg, 1988).

The central concept of the model is that different dimensions of experience acquire different degrees of *communicative significance*, or selective attention (Nosofsky, 1986). This is true of both form and meaning, which are represented in the same psychological space. At the beginning of learning, all dimensions are equally, and weakly, weighted, and the model responds in a graded fashion, on the basis of overall similarity. Later in learning, however, some dimensions become very significant, and others insignificant. The model then responds categorically, in the all-or-none fashion characteristic of symbolic representations. It is this transition that we suggest underlies the emergence of words, as symbols.

Formal presentation: Given any input, the model computes the distance of the current input from the weight vector of each lexicon node:

$$d_i = \sqrt{\sum_j s_j (i_j - w_{ij})^2}$$

Here s_j is the communicative significance of dimension j, i_j is the current value (+/-1) of input dimension j, and w_{ij} is the weight on the connection from input node j to lexicon node i. Note that distance is computed over both

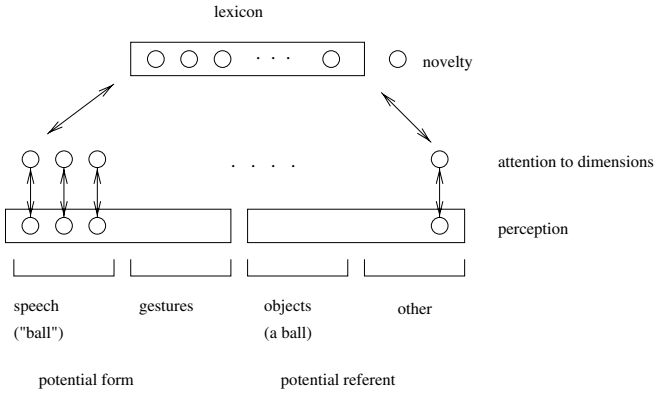


Figure 1: A model of early word learning.

form and meaning dimensions, together. The activation of the lexicon node is then an exponential function of this psychological distance (Shepard, 1987):

$$a_i = exp(-d_i)$$

Thus, the activation for a lexicon node i will be at its maximum (1.0) when d_i is zero - which will occur when the input and weight vectors are identical along significant dimensions. There is also a "novelty node", a novelty detector that is activated to the extent that no existing lexicon node is activated:

$$a_{novelty} = 1 - max_i(a_i)$$

Given these activations, we can compute a probability distribution over the lexicon, including the novelty node, using the Luce choice rule:

$$p_i = \frac{a_i}{\sum_j a_j}$$

A new lexicon node is then created with probability $p_{novelty}$ – the probability associated with the novelty node. If this node-creation does occur, the newly-created node k is incorporated into the lexicon, and its weights are set by the current input values, multiplied by $p_{novelty}$ (Hebb, 1949). Thus, the more clearly novel the input is, the stronger the weights on the newly allocated nodes will be.[1]

$$w_{kj} = i_j \times p_{novelty}$$

If a new node is not created, the most highly activated node k in the lexicon is selected for training. The model is then trained under gradient descent, with a target value of 1.0 for p_k, and target values of 0.0 for all other $p_i, i = k$. This will reinforce node k's representation of the form-meaning pairing currently presented as input, moving its weights w_k closer to the current input, and those of competing nodes farther away. Similarly, the s values (significance weights) will be adjusted to help discriminate lexicon nodes from each other. We then train again with only the form as input, and with the same target outputs. And finally, we train again with only the referent as input, and with the same target outputs. This three-step training causes the selected node k to act as a category node for both form and referent - and thereby to link the two.

The equations above, with the exception of the novelty node activation and creation, are adapted from existing models of categorization (Kruschke, 1992; Nosofsky, 1986). Merriman (1999) has shown that a similar formalism can account for the mutual exclusivity bias and shape bias in word-learning – the present model builds on this work. The combination of form and meaning in a single representation echoes the linguistic production model of Dell et al. (1997). This links theoretically central aspects of our model to existing models of related processes, models that have already received empirical support in their own right.

Testing: Once a set of words has been learned, one may test the model on either comprehension or production of a learned word. In a comprehension test, the word form is supplied to the network, but no referent is supplied. A winning lexicon node k is selected, as above, but from a competition based on the form dimensions only. The referent dimensions of node k's weight vector w_k are then projected down to the referent inputs:

$$i_j = w_{kj}$$

This reconstructed referent constitutes the model's response to the word form supplied as input. Production tests proceed analogously, but with the referent supplied as input, and the form produced as output.

Accounting for Existing Data

In the simulations reported here, the model was trained on a dataset of form-meaning pairings, and tested at various points during training. There were 6 dimensions for form: 4 were significant (such that a pattern over these dimensions was predictive of the referent), and 2

[1] Nodes that are created but not revisited within a given number of epochs are pruned from the lexicon. In current simulations, that number of epochs is 1.

were insignificant (not predictive). Similarly, there were 6 dimensions for meaning: 4 significant (such that a pattern over these dimensions was predictive of the form), and 2 insignificant (not predictive). The training set consisted of 75 variants of 5 words; a variant of a word preserved the significant dimensions of the word while altering insignificant dimensions. The words and their variants were represented by either +1 or -1 on each dimension. The specific values were chosen randomly, subject to the constraint that patterns over the significant dimensions of form be predictive of significant dimensions of meaning, and vice versa. The model was trained on this dataset for 100 epochs. The learning rate was 0.05. As expected, the significance weights differentiated significant from insignificant dimensions increasingly clearly over the course of training. At epoch 6, the difference between the average significance weight over dimensions intended to be significant and the average significance weight over dimensions intended to be insignificant was 1.06608. By epoch 60, this difference was 2.74704, reflecting clearer differentiation with training.

During training, tests were performed in order to probe the model's behavior on the four empirical trends noted at the beginning of the paper. In all cases, this involved presenting a new form-meaning pairing for the model to learn, and then determining the probability of correct comprehension or production. We define a "correct" response to be one that is within 0.9 of the target along all significant dimensions (which generally vary from -1 to 1), using the comprehension and production output rules outlined above. We then calculate the summed probability across all lexicon nodes that would produce a correct response: this yields the probability of correct response.

Ease of learning: We first tested how easily the model could learn a novel form for a novel object, and how that ease changed with age. A new word was created that was representationally distant from the existing words in the training set – specifically, each of form and meaning in this new word differed from those for existing words along all significant dimensions. We shall refer to this word henceforth as the "novel" word. We examined the probability of correct comprehension of this new word after simulating one learning trial on the word at two points during the learning of the training set mentioned above: after 6 epochs, and after 60 epochs. The results are displayed in Figure 2(a). As the model's space stretches along significant dimensions, this new word is increasingly easily learned, eventually being reliably correctly comprehended given only 1 training trial. This progression into 1-trial learning reflects the behavior of 1-2 year old children. Importantly, once the model had learned the novel word, it was removed from the lexicon; thus, later "ages" of the model did not have the benefit of earlier training on the novel word – only of an appropriately stretched psychological space, which caused the word to be perceived as distinct, and therefore easily remembered.

Honing of linguistic form: We next examined the learning of a new word that was *similar in form* to an existing word in the training set. The form of this new word differed from that of an existing word in the lexicon by only 1 significant dimension, while the meaning dimensions differed from other words along all significant dimensions. Thus, this test simulates the potential confusion of learning "bih" and "dih" associated with different sorts of objects (Stager and Werker, 1997). The probability of correct comprehension after one training trial is shown by the crosshatched bars in Figure 2(b). In this figure, the solid bars duplicate the presentation of the model's behavior on the novel (dissimilar) word in (a), for purposes of comparison. As is true of children, similar words are initially somewhat more difficult to learn than are globally dissimilar words. However, eventually these similar words are also successfully learned given one training trial, as the relevant dimensions of space are highlighted, counteracting the confusing similarity. This allows fast learning of minimal pairs such as "bit" and "pit".

Honing of meaning: We were interested in determining whether the model would exhibit a strengthening shape bias, as children do. To test this, as before, we trained the model on the novel word, and then tested the probability of producing the novel word for a different object, which differed from the original in meaning along insignificant (non-shape) dimensions only. This probability of generalization is shown in Figure 2(c). The increasing strength of generalization along the significant dimensions follows directly from the increasing perceived communicative significance of those dimensions (see Merriman (1999) for a similar demonstration). This is analogous to the honing of linguistic form. This account is incomplete however. In actuality the shape bias applies only to object names; thus an additional mechanism would be required to determine whether a given word is an object name, and therefore whether the model's bias should apply.

Synonyms: A synonym for an existing word in the lexicon should be difficult to learn, since it is similar (identical in meaning) to that existing word. Figure 2(d) shows that this is the case in the present model; Merriman (1999) reports similar results with a similar model. The probability of correct comprehension after one training exposure is initially lower for a synonym of an existing word in the lexicon than it is for the (dissimilar) novel word examined above. This matches the findings of Liittschwager and Markman (1994). Eventually however the synonym and the non-synonym are approximately equally likely to be learned – as is eventually true with children. In the model, this is accounted for by the stretching of the underlying psychological space with age, such that even similar lexical entries are kept distinct, and thereby effectively learned.

Predictions

The model makes two predictions. The first is that young children should experience difficulty learning *homonyms* (a single form with multiple meanings, such as the "bank" of a river, and a "bank" as a financial institution). Moreoever, this difficulty should be correlated

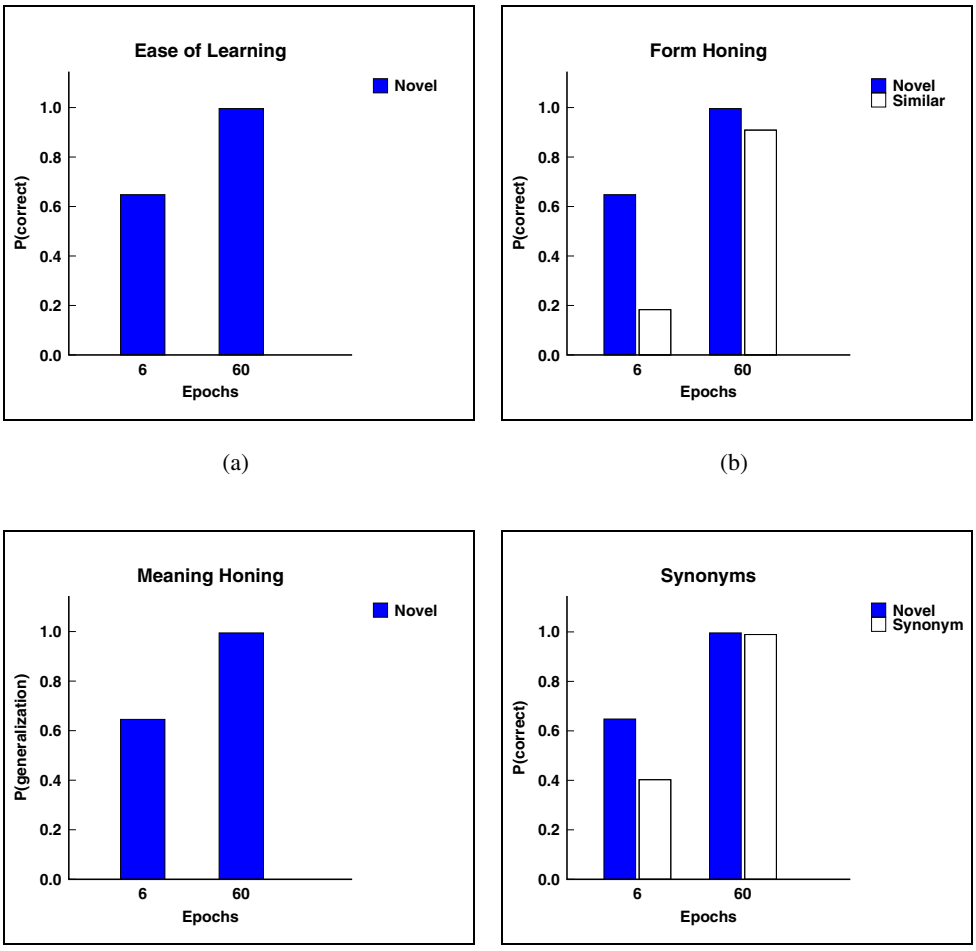


Figure 2: Developmental trends, as exhibited by the model.

with the difficulty of learning synonyms, as the reason for the difficulty is analogous. Half of the model's lexical representation for a homonymous or synonymous word will be identical to that for another word in the lexicon: the identical half is the form for homonyms, and the meaning for synonyms. This means that the two lexical entries in question will tend to be nearer each other in psychological space than two non-homonyms or two non-synonyms, and will therefore interfere with each other. Doherty (2000) has found that understanding of homonymy is strongly associated with understanding of synonymy, in 3-4 year old children.[2] Similar tests on younger children would more directly test this prediction.

The second prediction also concerns the interaction of form and meaning. As we have seen, 14-month-olds have trouble associating similar sounds such as "bih" and "dih" with different referents (Stager and Werker, 1997). On the model's account, this is because the forms are too similar, such that the two lexical representations lie confusingly near each other in psychological space. But since that space contains both form and meaning dimensions, the model predicts that an exaggerated *semantic* difference between the referents should compensate for the confusing formal similarity of "bih" and "dih" in such a task, and should make learning easier. This prediction has not yet been tested.

Discussion

1- to 2-year old children seem to undergo a qualitative change in the manner in which they learn words. It has been suggested that this behavioral change reflects a conceptual insight into the symbolic nature of words. The model we have presented, however, suggests a different, and more parsimonious, account of the same phenomenon. The behavioral change may result not from

[2]These were both also correlated with understanding of false belief.

an abrupt insight, but rather from an associative learner gradually determining which aspects of the world are relevant for communication. In this manner, the symbolic use of words may emerge from an associative base.

Acknowledgments

This work was supported by grant R01-DC03384 from the National Institutes of Health.

References

Baldwin, D., Markman, E., Bill, B., Desjardins, R., and Irwin, J. (1996). Infants' reliance on a social criterion for establishing word-object relations. *Child Development*, 67:3135–3153.

Bird, E. K. and Chapman, R. S. (1998). Partial representation and phonological selectivity in the comprehension of 13- to 16-month-olds. *First Language*, 18:105–127.

Carpenter, G. A. and Grossberg, S. (1988). The ART of adaptive pattern recognition by a self-organizing neural network. *Computer*, pages 77–88.

Corkum, V. and Moore, C. (1998). The origins of joint visual attention in infants. *Developmental Psychology*, 34:28–38.

Cottrell, G. and Plunkett, K. (1994). Acquiring the mapping from meaning to sounds. *Connection Science*, 6(4):379–412.

Dell, G., Schwartz, M., Martin, N., Saffran, E., and Gagnon, D. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, 104(4):801–838.

Doherty, M. J. (2000). Children's understanding of homonymy: Metalinguistic awareness and false belief. *Journal of Child Language*, 27:367–392.

Dromi, E. (1987). *Early Lexical Devlopment*. Cambridge University Press, New York.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. MIT Press, Cambridge, MA.

Gershkoff-Stowe, L. and Smith, L. B. (1997). A curvilinear trend in naming errors as a function of early vocabulary growth. *Cognitive Psychology*, 34:37–71.

Hebb, D. (1949). *The Organization of Behavior*. Wiley.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99:22–44.

Liittschwager, J. and Markman, E. (1994). Sixteen and 24-month-olds' use of mutual exclusivity as a default assumption in second label learning. *Developmental Psychology*, 30:955–968.

McShane, J. (1979). The development of naming. *Linguistics*, 17:879–905.

Merriman, W. (1999). Competition, attention, and young children's lexical processing. In MacWhinney, B., editor, *The Emergence of Language*, pages 331–358. Lawrence Erlbaum Associates, Mahwah, NJ.

Namy, L. L. and Waxman, S. R. (1998). Words and gestures: Infants' interpretations of different forms of symbolic reference. *Child Development*, 69:295–308.

Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39–57.

Plunkett, K., Sinha, C., Moller, M., and Strandsby, O. (1992). Symbol grounding or the emergence of symbols? vocabulary growth in children and a connectionist net. *Connection Science*, 4:293–312.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, 237:1317–1323.

Smith, L. B., Jones, S., Gershkoff-Stowe, L., and Samuelson, S. (1999). The origins of the shape bias. Submitted to the *Monographs of the Society for Research in Child Development*.

Stager, C. L. and Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, 388:381–382.

Woodward, A. L. and Hoyne, K. L. (1999). Infants' learning about words and sounds in relation to objects. *Child Development*, 70:65–77.

A Knowledge-Resonance (KRES) Model of Category Learning

Bob Rehder (bob.rehder@nyu.edu)
Department of Psychology, New York University, 6 Washington Place
New York, NY 10003 USA

Gregory L. Murphy (gmurphy@s.psych.uiuc.edu)
Beckman Institute, University of Illinois, 405 N. Mathews Ave
Urbana, IL 61801 USA

Abstract

In this article we present a connectionist model of category learning that takes into account the prior knowledge that people bring to many new learning situations. This model, which we call the *Knowledge-Resonance Model* or *KRES*, employs a recurrent network with bidirectional connections which are updated according to a *contrastive-Hebbian learning rule*. When prior knowledge is incorporated into a KRES network, the KRES activation dynamics and learning procedure accounts for a range of empirical results regarding the effects prior knowledge on category learning, including the accelerated learning that occurs in the presence of knowledge, the reinterpretation of features in light error correcting feedback, and the unlearning of prior knowledge which is inappropriate for a particular category.

A traditional assumption in category learning research is that learning is based on those category members people observe and is relatively independent of the prior knowledge that they already possess. According to this *data-driven* or *empirical learning* view of category learning, people associate observed exemplars and the features they display (or a summary representation of those features such as a prototype or a rule) to the name or label of the category. In this account there is neither need nor room for the influence of the learner's prior knowledge of how those features are related to each other or other concepts on the learning process. In contrast, the last several years has seen a series of empirical studies that demonstrate the dramatic influence that a learner's prior knowledge often has on the learning process in interpreting and relating a category's features to one another and other concepts. Indeed, knowledge effects have been demonstrated in every area of conceptual processing in which they have been investigated (see Murphy, 1993, for a review).

The goal of this article is to introduce a theory of category learning that accounts for the effects of prior knowledge on the learning of new categories. This theory, which we refer to as the *Knowledge-Resonance Model*, or *KRES*, is a connectionist network that specifies prior knowledge in the form of existing concepts and relations between concepts. We will show that when knowledge is incorporated into a KRES network, KRES's activation and learning procedures account for a number of empirical results regarding the effects of prior knowledge on category learning.

Other connectionist models have been proposed to account for the learning of new categories (e.g., Gluck & Bower, 1988; Kruschke, 1992), and these models have generally used feedforward networks (i.e., activation flows only from inputs to outputs) and learning rules based on error signals that traverse the network from outputs to inputs (e.g., backpropagation). KRES departs from these previous models in two regards. First, rather than feedforward networks, KRES uses *recurrent networks* in which activation is allowed to flow not only from inputs to outputs but also from outputs to inputs and back again. Recurrent networks respond to inputs by each unit iteratively adjusting its activation in light of all other units until the network "settles," that is, until change in units' activation levels ceases. This settling process can be understood as an interpretation of the input in light of the knowledge encoded in the network. As applied to the categorization problems considered here, a KRES network accepts inputs that represent an object's features, and interprets (i.e., classifies) that object by settling into a state in which the object's correct category label is active.

Second, rather than backpropagation, KRES employs *contrastive Hebbian learning* (CHL) as a learning rule (Movellan, 1989; O'Reilly, 1996). Backpropagation has been criticized for being neurally implausible because it assumes nonlocal information regarding the error generated from corrective feedback in order for connection weights to be updated. In contrast, CHL transmits error by using the same connections that propagate activation. During an initial *minus phase*, a network is allowed to settle in light of an input pattern. In the ensuing *plus phase*, the network is provided with what serves as error-corrective feedback by being presented with the output pattern that should have been computed during the minus phase and allowed to resettle in light of that (correct) output pattern. Connection weights are then updated as a function of the difference between the activation of units in the two phases.

In the following sections we first describe KRES and then present three simulations of human category learning data. We will show how KRES's successes can be attributed to its recurrent network that allows category features to be interpreted in light of prior knowledge, and the CHL learning algorithm that allows (re)learning of all connections in a network, including those that represent prior knowledge.

The Knowledge-Resonance Model (KRES)

An example of a KRES model is presented in Figure 1. In Figure 1, circles depict *units* that represent concepts that are

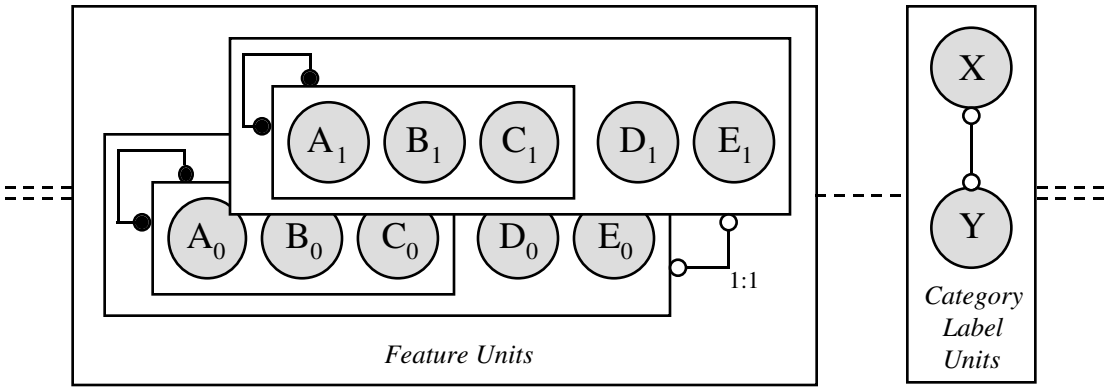


Figure 1. A sample KRES model.

either category labels (X and Y) or features (A_0, A_1, B_0, B_1, C_0, C_1, etc.). To simplify the depiction of connections among units, units are organized into *layers* specified by rectangles. Solid lines between layers represent connections among units. Solid lines terminated with black circles are excitatory connections, whereas those terminated with hollow circles are inhibitory connection. Dashed lines represent new, to-be-learned connections. Two connected layers are fully connected (i.e., every unit is connected to every other unit), unless annotated with "1:1" (i.e. "one-to-one") in which case a unit in a layer is connected to only one unit in the other layer. Finally, double dashed lines represent sources of external inputs. As described below, both the feature units and the category label units receive external input, although at different phases of the learning process.

We now describe the basic elements of KRES, which include its representation assumptions, activation dynamics (i.e., constraint satisfaction), and learning via CHL.

Representational Assumptions

At any time a unit has a level of activation in the range 0 to 1 that represents the activation of the concept. A unit *i*'s activation act_i is a sigmoid function of its total input,

$$act_i = 1 / [1+ \exp (-total\text{-}input_i)]$$

and its total input comes from three sources,

$$total\text{-}input_i = net\text{-}input_i + external\text{-}input_i + bias_i.$$

Network input represents the input received from other units. External input represents the presence of (evidence for) the concept in the external environment. Finally, a unit's bias can be interpreted as a measure of the prior probability that the concept is present in the environment.

In many applications, two or more features might be treated as mutually exclusive values on a single dimension. In Figure 1 the stimulus space is assumed to consist of five binary valued dimensions, with A_0 and A_1 representing the values on dimension A, B_0 and B_1 the values on dimension B, etc. To represent that these feature pairs are mutually exclusive they are linked by inhibitory connections. The category labels X and Y are also assumed to be mutually exclusive and are linked by an inhibitory connection.

Connections between units are symmetric, that is, $weight_{ij} = weight_{ji}$. A unit's network input is computed by multiplying the activation of each unit to which it is connected by the connection's weight, and then summing over those units in the usual manner,

$$net\text{-}input_i = \sum_j act_i * weight_{ij} .$$

KRES primarily represents prior knowledge in the form of prior relations between features. For example, in Figure 1 it is assumed that features A_0, B_0, and C_0 are related by prior knowledge, as are features A_1, B_1, and C_1. These relations are rendered as excitatory connections between the features. In KRES prior knowledge can also be represented in the form of preexisting concepts (i.e., units) and excitatory connections that link those preexisting concepts to the feature units (see Simulation 3 below).

Classification via Constraint Satisfaction

Before a KRES model is presented with input that represents an object's features, the activation of each unit is initialized to a value determined solely by its bias. The external input of a feature unit is then set to 1 if the feature is present in the input, -1 if it is absent, and 0 if its presence or absence is unknown. The external input of all other units is set to 0. The model then undergoes a multi-cycle constraint satisfaction processes which involves updating the activation of each unit in each cycle in light of its external input, its bias, and its current network input. (In each cycle, the serial order of updating units is determined by randomly sampling units without replacement.) After each cycle the *harmony* of the network is computed, given by,

$$harmony = \sum_i \sum_j act_i * act_j * weight_{ij} . \qquad (1)$$

Constraint satisfaction continues until the network settles, as indicated by a change in harmony from one cycle to the next of less than 0.00001.

The activation of units X and Y that result from this settling process represent the evidence that the current input should be classified as an X or Y, respectively. These activation values can be mapped into a categorization decision in the standard way, that is, according to Luce's choice axiom,

$$choice\text{-}probability\ (X, Y) = act_X \ / \ (act_X + act_Y) .$$

Contrastive Hebbian Learning (CHL)

As described earlier, the settling of a network that results from presenting just the feature units with input is referred to as the minus-phase. In the plus-phase, error-correcting feedback is provided to the network by setting the external

inputs of the correct and incorrect category label units to 1 and –1, respectively, and allowing the network to resettle in light of these additional inputs. We refer to the activation values of unit i that obtain after the minus and plus phases as act_i^- and act_i^+, respectively. After the plus phase the connection weights are updated according to the rule,

$$\Delta weight_{ij} = lrate * (act_i^+ * act_j^+ - act_i^- * act_j^-) \qquad (2)$$

where *lrate* is a learning rate parameter.

Network Training

Before training a KRES network, all connections weights are set to their initial values. In the following simulations, all to-be-learned connections are initialized to a random value in the range [-0.1, 0.1], and the biases of all units are initialized to 0. As in the behavioral experiments we simulate, training consists of repeatedly presenting a set of training patterns in blocks with the order of the patterns randomized within block. Training continues until the error for a block falls below an error criterion of 0.10. The error for a block is computed by summing the errors associated with each training pattern in the block and dividing by the number of patterns. The error associated with a training pattern is the sum of the squared differences between the activation levels of the category label units and their correct values (0 or 1).

KRES Simulation of Empirical Data

We present KRES simulations of three empirical data sets that illustrate the effect of prior knowledge on category learning. The KRES model was rerun ten times with a different set of random weights, and the results reported below are averaged over those ten runs.

Simulation 1: Murphy and Allopenna (1994)

In the literature on category learning with prior knowledge, perhaps the most pervasive effect is that learning is dramatically accelerated when the prior knowledge is consistent with the empirical structure of training exemplars. For example, Murphy and Allopenna (1994, Experiment 2), presented examples of two categories the features of which either could (Theme Condition) or could not (No Theme Condition) be related to one another. In the Theme condition one category had six typical features that could be related because they could be construed as features of arctic vehicles ("drives on glaciers," "made in Norway," "heavily insulated," etc.) whereas the other category had six typical features that could be construed as features of jungle vehicles ("drives in jungles," "made in Africa," "lightly insulated," etc.). In the No Theme condition, the typical features of the categories could not be related to one another. Exemplars also possessed three knowledge-irrelevant features which were not predictive of category membership. Murphy and Allopenna found that participants reached a learning criterion in fewer blocks in the Theme (2.5) versus the No Theme condition (4.1), a result the authors attribute to the knowledge relating the features in the Theme condition.

This experiment was simulated by a KRES model like the one shown in Figure 1 with 18 features representing the two values on 9 binary dimensions. In the Theme but not the No Theme condition the six related features in each of the two

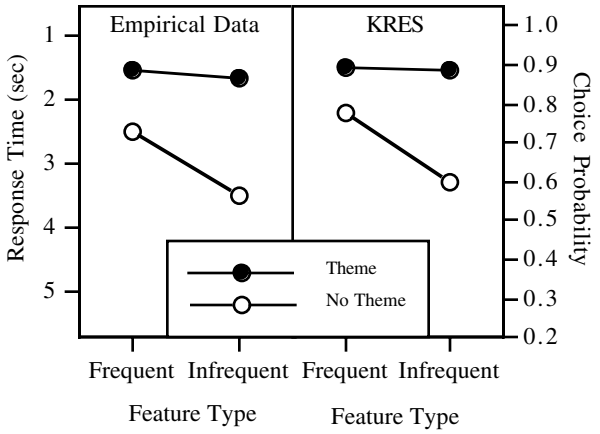


Figure 2. Results from Murphy & Allopenna (1994).

categories were linked with excitatory connections. The weight on these excitatory connections was initialized to 0.4, the inhibitory connections were initialized to –2.0, and the learning rate was set to 0.10.

The results indicated that KRES reproduces the learning advantage found in the Theme condition: The error criterion was reached in fewer blocks as compared to the No Theme condition (2.0 vs. 4.0). This advantage can be attributed to KRES's use of recurrent networks: The mutual excitation of knowledge-relevant features in the Theme condition resulted in higher activation values for those units, which in turn led to the faster growth of the connection weights between the features and category label units (according to the CHL learning rule Eq. 2). Once some learning of those connections has occurred, the higher activation of the features also leads to greater activation of the category labels themselves.

Murphy and Allopenna also varied the frequency with which the six knowledge-relevant features appeared during training, and then tested how subjects classified those features during an ensuing test phase. The left side of Figure 2 indicates that, as expected, RTs for these single-feature classification trials were shorter for frequent versus infrequent features in the No Theme condition. In contrast, in the Theme condition RTs were insensitive to features' empirical frequency. This pattern of results was also reflected in subjects' categorization accuracy. (Note Figure 2's RT scale has been inverted to facilitate comparison with KRES's choice probabilities presented below.)

To determine whether KRES would also exhibit these effects, after training the model was presented with single features. That is, the unit representing that feature was given an external input of 1, the unit representing the other feature on the same dimension was given an input of –1, and all other units were given an input of 0. The right side of Figure 2 indicates that KRES's choice probabilities reproduce the pattern of results for the single-feature tests. In KRES, infrequently presented knowledge-relevant features are classified nearly as accurately as frequently presented ones because during training those features were activated by inter-feature excitatory connections even on trials in which they were not presented, and hence were associated with the category label nearly as strongly as knowledge-relevant features that were frequently presented.

Simulation 2: Kaplan and Murphy (2000)

Simulation 1 provides evidence in favor of KRES's use of recurrent networks to accelerate learning by amplifying the activation of features interconnected by prior knowledge. However, another distinctive characteristic of KRES is that the category label units are also recurrently connected to the features. In this section we provide evidence that activation also flows backwards from category label units.

Using a modified version of the materials used in Murphy and Allopenna (1994), Kaplan and Murphy (2000, Experiment 4) provided an especially dramatic demonstration of the effect of prior knowledge. In that study, participants were presented with training examples that contained only *one* of the knowledge-relevant features and up to six knowledge-irrelevant features that were predictive of category membership. That is, the single knowledge-relevant feature in each exemplar had prior associations only to features in other category exemplars. Under these conditions, one might have predicted that participants would be unlikely to notice the relations among the features in different exemplars, especially given that those features were each embedded in an exemplar with many knowledge-irrelevant features. In fact, participants in this Intact Theme condition reached a learning criterion in fewer blocks (2.7) than did those in a No Theme condition (5.0) in which the categories had the same empirical structure but no relations among features.

We simulated this experiment with a KRES model with 22 features on 11 binary dimensions. In the Intact Theme condition the features within the two sets of six knowledge-relevant features were inter-related with excitatory connections, as in Simulation 1. The weight on these excitatory connections was initialized to 0.35, the inhibitory connections were set to –2.0, and the learning rate was set to 0.10.

KRES reproduced the learning advantage found in the Intact Theme condition (3.0 blocks) as compared to the No Theme condition (5.4). This advantage obtained because even though each training pattern in the Intact Theme condition contained only one knowledge-relevant feature, that feature tended to activate the knowledge-relevant features to which it was connected, and hence the connections between each knowledge-relevant feature and its correct category label were strengthened on every trial to at least some degree.

After each training block, Kaplan and Murphy also presented test blocks in which participants classified each of the 22 features. The left side of Figure 3 indicates that as expected after the final block of training participants in the No Theme condition were faster at classifying those features that appeared in several training exemplars (Characteristic features) than those that appeared in just one (Idiosyncratic features). In contrast, in the Intact Theme condition participants were faster at classifying the Idiosyncratic features, because they were also knowledge-relevant. Unexpectedly, Intact Theme participants were also faster at classifying the Characteristic features (i.e., the knowledge-irrelevant features) even though those features were not related via prior knowledge, and even though Intact Theme participants had experienced fewer training blocks on average (2.7 vs. 5.0).

This latter result is a challenge for many standard connectionist accounts of learning, because in such accounts the better learning associated with knowledge-relevant features would be expected to *overshadow* the learning of knowledge-irrelevant features (Gluck & Bower, 1988)–that is, these features should be worse with knowledge than without as a result of them competing with the stronger knowledge-relevant features. In contrast, Figure 3 indicates that KRES is able to account for the better learning (or in some experiments, equal learning) of the knowledge-irrelevant features in the Intact Theme condition. This result can be attributed to the use of recurrent connections to the category label units. After some excitatory connections between the knowledge-irrelevant features and category labels have been formed, the knowledge-relevant and -irrelevant features began to activate each other through the category node. This greater activation of the knowledge-irrelevant features leads to accelerated learning of their connection weights to the category labels. That is, KRES's use of recurrent networks compensates for the effects of cue competition found in the usual feedforward network.

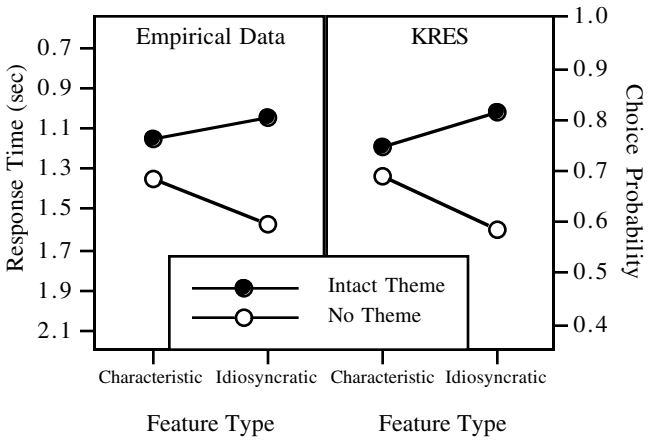


Figure 3. Results from Kaplan & Murphy (2000). In the Intact Theme condition Idiosyncratic features are knowledge-relevant and Characteristic features are knowledge-irrelevant.

Simulation 3: Wisniewski and Medin (1994)

In a final simulation we demonstrate the efficacy of contrastive-Hebbian learning to update weights on connections not involving the category label units. In particular, we examine KRES's ability to update connections representing prior knowledge that is inappropriate in the current context.

Wisniewski and Medin (1994, Experiment 2) present empirical results that call into question the assumption of standard theories of category learning that features can be identified prior to learning. Participants were shown two categories of line drawings of persons that were described as drawn by *creative* and *non-creative children* or by *farm* and *city kids*. Wisniewski and Medin chose to use line drawings to illustrate that what constitutes a feature in a stimulus depends on the prior expectations that one has about its possible category membership. For example, they found that participants would assume the presence of *abstract features* about a category depending on the category's label (e.g., creative children's drawings depict unusual amounts of detail and characters performing actions) and examine the drawings for concrete evidence of those abstract features in order to

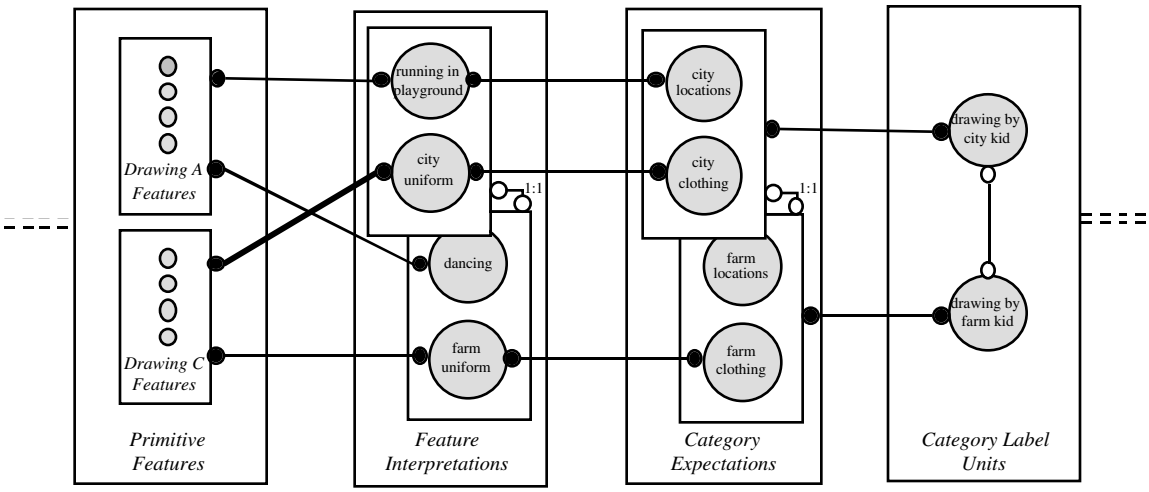


Figure 4. KRES model for Simulation 3.

determine its category membership. They also found that the feedback participants received about category membership led them to change their original interpretation of certain features of the line drawings. For example, after first interpreting a character's clothing as a farm "uniform" (and categorizing the picture as drawn by a farm kid), some participants reinterpreted the clothing as a city uniform after receiving feedback that the picture was drawn by a city kid.

To demonstrate these effects with KRES, we imagined a simplified version of the materials of Wisniewski and Medin's in which there were only two drawings. One drawing (Drawing A), was of a character performing an action interpretable either as climbing in a playground or dancing. In the other (Drawing C), a character's clothing could be seen as a farm uniform or a city uniform. These alternative interpretations are represented in the left side of the KRES model of Figure 4. Whereas we assume the two interpretations of Drawing A are equally likely, we assume that a city uniform is the more likely interpretation of Drawing C (as depicted by the heavier line connecting the features of Drawing C and their city uniform interpretation). The alternative interpretations are connected with inhibitory connections representing that only one interpretation is correct.

The model of Figure 4 was presented with the problem of learning to classify Drawing A as done by a city kid, and Drawing C by a farm kid. We represented the expectations or hypotheses that Wisniewski and Medin found that learners form in the presence of meaningful category labels such as *farm* or *city kids* as units connected via excitatory connections to the category labels, as shown in the right side of Figure 4. In Figure 4, city and farm kids are expected to be in locations and wear clothing appropriate to cities and farms. These expectations are in turn related by excitatory connections to the picture interpretations that instantiate them: climbing in a playground instantiates a city location, and city and farm uniforms instantiate city and farm clothing, respectively. In Figure 4, all inhibitory connections were set to –3.0 and all excitatory connections were set to 0.25, except for those between Drawing C's features and their city uniform interpretation, which were set to 0.30.

Before a single training trial is conducted, the prior knowledge incorporated into this KRES model is able to decide on a classification of both drawings. Upon presentation of Drawing A, its two interpretations, climbing-in-a-playground or dancing are activated, and climbing-in-a-playground in turn activates the city location expectation, which in turn activates the category label for city kids' drawings. The drawing is correctly classified as having been drawn by a city kid. Moreover, as the network continues to settle, activation is sent back from the category label to the climbing-in-a-playground unit. As a result, the climbing-in-a-playground interpretation of Drawing A is more active than the dancing interpretation when the network settles. That is, the top-down knowledge provided to the network results in the resolution of an ambiguous feature (i.e., the action is interpreted as climbing in a playground rather than dancing). Wisniewski and Medin found that the same drawing would be interpreted as depicting dancing instead when participants were required to classify the drawings as having been done by creative or noncreative children.

Similarly, upon presentation of Drawing C, its two interpretations are activated, but because the city uniform interpretation receives more input as a result of its larger connection weight, it quickly dominates the farm uniform interpretation. As a result, the category label for city kids' drawings becomes active (via the city clothing expectation). That is, the drawing is *incorrectly* classified as having been drawn by a city kid. However, error feedback results in the model changing its interpretation of Drawing C. During the model's plus phase, the farm kids' category label is more active than the city kids' label as a result of the external inputs those units receive. The activation emanating from the farm kids' label leads to the activation of the farm clothing expectation and then the farm uniform feature interpretation, which ends up dominating the city uniform unit.

This result indicates that KRES can reinterpret features in light of error feedback. The more important question, however, is whether KRES can *learn* this new interpretation so that Picture C (or a similar picture) will be correctly classified in the future. The left side of Figure 5 shows the changes to the connection weights brought about by the CHL learning rule with a learning rate of 0.3 as a function

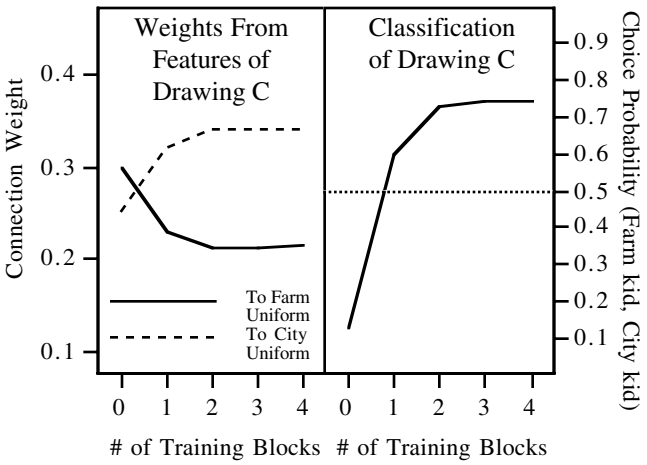


Figure 5. Results from Simulation 3.

of number of blocks of training on the two drawings. Figure 5 indicates that the connection weights associated with the interpretation of Drawing C as a city uniform rapidly decrease from their starting value of 0.30, while the weights associated with Drawing C's interpretation as a farm uniform increase from their starting value of 0.25. As a result, after just one training block KRES's classification of Drawing C switches from being done by a city kid to a farm kid (as indicated by the choice probabilities shown in the right side of Figure 5). That is, KRES uses the error feedback it receives to learn a new interpretation of Drawing C.

General Discussion

We have presented a new model of category learning that attempts to account for the influence of prior knowledge that learners often bring to the task of learning a new category. KRES utilizes a recurrent network in which knowledge is encoded in the form of connections among units. We have shown the changes brought about by this recurrently-connected knowledge to the interpretations and reinterpretations of a category's features provides a reasonable account of three data sets exhibiting the effects of prior knowledge on category learning. In Simulation 1 we demonstrated how KRES's recurrent network provides a pattern of activation among features that accounts for the finding that knowledge accelerates the learning of connections to category labels. In Simulation 2 we demonstrated that the presence of knowledge does not inhibit the learning of knowledge-irrelevant features, a striking result in light of well-known learning phenomena such as cue competition. In Simulation 3 top-down flow of activation was instrumental in KRES's success in resolving the ambiguity surrounding the interpretation of a perceptual features. Moreover, the CHL learning rule allowed the knowledge responsible for one interpretation of an ambiguous feature to be unlearned and a new interpretation learned when the network was provided with feedback regarding the stimulus's correct category.

KRES departs from previous connectionist models that attempt to account for the effects prior knowledge with feedforward networks. For example, Heit & Bott (2000) have proposed a model, *Baywatch*, that assumes that features send activation to prior concepts, that both the features and the prior concepts send activation to the category label units, and that learning consists of learning the connections to the category labels. Although we believe that existing categories often aid the learning of new categories (e.g., our knowledge of VCRs helps us understand DVD players), the Baywatch approach is limited to the learning of new categories that are essentially refinements of existing concepts. In contrast, KRES only assumes the presence of relations between features to account for the data in Simulations 1 and 2, and hence is able to learn truly new concepts, not just refinements of existing ones.

There remains much to be discovered about the properties of recurrent networks and contrastive Hebbian learning with regard to the learning of categories. However, we believe that recurrent networks are likely to be critical to any attempt at accounting for the effects of prior knowledge on category learning. For example, standard feedforward networks seem intrinsically unable to account for (a) the accelerated learning produced by prior knowledge without presupposing prior knowledge of the to-be-learned category, (b) the effects of top-down knowledge on resolving ambiguous features, and (c) the reinterpretation of ambiguous features in light of feedback regarding category membership.

Acknowledgements

This work was supported by NSF Grant SBR 97-20304.

References

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117*, 227-247.

Heit, E., & Bott, L. (2000). Knowledge selection in category learning. In D. L. Medin (Ed.), *The psychology of learning and motivation.* (pp. 163-199). Academic Press.

Kaplan, A. S., & Murphy, G. L. (2000). Category learning with minimal prior knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 829-846.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22-44.

Movellan, J. R. (1989). Contrastive Hebbian learning in the continuous Hopfield model. In D. S. Touretzky, G. E. Hinton, & T. J. Sejnowski (Eds.), *Proceedings of the 1989 Connectionist Models Summer School* .

Murphy, G. L. (1993). Theories and concept formation. In I. V. Mechelen, J. Hampton, R. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis.* (pp. 173-200). Academic Press.

Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 904-919.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation, 8*, 895-938.

Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science, 18*, 221-282.

Regularity and Irregularity in an Inflectionally Complex Language: Evidence from Polish

Agnieszka Reid (agnieszka.reid@mrc-cbu.cam.ac.uk)

William Marslen-Wilson (william.marslen-wilson@mrc-cbu.cam.ac.uk)
MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge CB2 2EF, UK.

Abstract

We report two experiments which address the question of whether there is support for single or dual mechanisms in the processing of Polish regular and irregular nouns and verbs. The data from an immediate cross-modal priming experiment (verbs and nouns) revealed that there was significant priming, irrespective of regularity, for both verbs and nouns. The results for verbs were replicated in a second experiment using delayed repetition auditory-auditory priming. The outcome from both experiments is in line with the results reported from research on Italian and French but not with English and German, and suggest that Polish regulars and irregulars are processed by a single underlying cognitive system.

Introduction

There has been a long lasting debate as to the implications of English regular and irregular verbs for the universal properties of human cognitive mechanisms (Marslen-Wilson & Tyler, 1998; Pinker, 1999). The protagonists of a dual mechanism approach (Pinker & Prince, 1991) have claimed that English regular past tense forms are processed by a symbolic rule, which concatenates allomorphic suffixes (*-ed* or *-d* or *-t*) to the verb stem, while the irregular past tense forms are processed as whole forms in an associative memory. In contrast, the proponents of the single mechanism (Plunkett & Marchman, 1993; Rumelhart & McClelland, 1986) claim that regular and irregular English verbs are processed and represented in the same way, by a distributed connectionist system, operating without symbols and syntax.

We focus here on the experimental work probing the online processing and representation of regular and irregular forms in the adult mental lexicon. Much of this work has of course focused on English, chiefly using a variety of priming paradigms. In early work, for example, Stanners, Neiser, Hernon and Hallet (1979) tested the representation of regular and irregular verbs using visual priming with a delay. The authors found that regular inflected verbs primed their bases as effectively as identity primes. In contrast the irregular past tenses primed their bases significantly less than identity primes. Marslen-Wilson, Hare and Older (1995) investigated the representation of English regular and irregular verbs using a cross-modal lexical decision paradigm. The authors contrasted regular verbs (*jumped/jump*), with irregular verbs which fell into two classes: semi-weak verbs (*burnt/burn*) with irregular alveolar inflection, and vowel change verbs (*gave/give*). The results showed that there was significant priming for the regulars, but no priming for either sub-class of irregular verbs.

Evidence for behavioural differences of this type, suggesting asymmetries in the underlying systems supporting the processing of regular and irregular forms, can be interpreted as support for a dual mechanism approach. It is important, however, to apply these techniques cross-linguistically, to develop a broader perspective on the processing consequences of regularity and irregularity, and to evaluate the universality of the claims being made.

Cross-linguistic behavioural studies

Although cross-linguistic research is in its early stages it has already suggested interesting contrasts between English and German as opposed to Italian and French.

Sonnenstuhl, Eisenbeiss and Clahsen (1999) probed the representation of regular and irregular German verbs as well as regular and irregular noun plurals, using the cross-modal paradigm. Significant priming for both regulars and irregulars was obtained. The regulars were as effective as identity primes, whereas the irregulars were significantly less effective than identity primes. The authors argued that these data, consistent with English, provide support for the dual mechanism where the regulars are processed as a stem and an affix, generated by a rule, whereas irregulars are processed as full lexical items.

Both English and German belong to the same West Germanic language family and share many similarities. In additional research we have begun to look at languages from different sub-families which show different distributions of regulars and irregulars as well as of inflectional suffixation.

Significant cross-modal priming has been found for Italian regular and irregular verbs (Orsolini & Marslen-Wilson 1997). For regular verb pairs, the prime was typically a first or third conjugation past definite form, paired with either an infintive target (as in ***gioc****-a-rono/****gioc****-are* 'they played/to play') or with a past participle target (as in ***am****-a-rono/****am****-a-to* 'they loved/loved'). These conditions were contrasted with irregular verb pairs, such as *sce-s-ero/scend-ere* 'they got down, past definite'/to get down, infinitive', with idiosyncratic alternations, comparable to English irregular alternations, but occurring in a much more constrained and predictable morphological environment. Significant priming

was found across the board, with no differences between conditions in the magnitude of priming. There was no evidence here, or in an accompanying study using elicitation techniques, to suggest underlying differences in the processing and representation mechanisms evoked by regular and irregular forms.

Comparable results were obtained for French, where Meunier and Marslen-Wilson (2000) investigated the processing of French verbs with regular and irregular alternations in both cross-modal priming and masked priming experiments. They distinguished three levels of irregularity, ranging from phonologically triggered alternations such as *sème/semer* 'I sow/to sow', through sub-regularities such as *teignent/teindre* 'they dye/to dye', to highly idiosyncratic alternations such as *iront/aller* 'they will go/to go. As in Italian, these irregularities are linguistically embedded in extensive, primarily regular inflectional paradigms. Again, strong priming, without significant differences in the effects for regulars and irregulars, was found across the board. These results, similar to those from Italian, provide no support for the view that different mechanisms handle the processing of regulars and irregulars.

The aim of this paper is to extend these cross-linguistic investigations to Polish, a representative of the Slavonic language family, where very little behavioural research has been conducted. A striking characteristic of Polish is the great richness of its inflectional systems and the wide distribution of phonological and morphophonological alternations, varying considerably in regularity. In addition, note that in Polish, as in French and Italian, but in contrast with English, both regular and irregular verbs are inflected by concatenation of an inflectional suffix with a stem. In English, this is true for regular past tenses, e.g. *jump/jump-ed*, but not for irregular past tense forms, e.g. *give/gave*. The distinction between Polish regular and irregular forms is made on the basis of a vowel or consonant change (an alternation) in inflected forms. Regular inflected forms have morphophonological alternations which are predictable from Polish morphophonology. Irregular forms have morphophonological alternations which are idiosyncratic and are not predictable from the rules of Polish morphophonology.

We report two preliminary studies which probe the processing of Polish regular and irregular verbs and nouns (Experiment 1) and verbs (Experiment 2). Before moving on to the experiments, a description of the relevant characteristics of Polish is necessary.

Polish inflectional morphology and alternations

Almost every word in Polish exists within a very extensive morphological paradigm: declensional for nouns, adjectives, numerals and pronouns and conjugational for verbs. Every verb in Polish is inflected according to one of the three conjugational paradigms (Laskowski, 1998). The basis of the division of verbal themes into conjugations is their ability to concatenate with the complex morpheme of the present tense. The themes which are members of conjugation I take in the present tense the endings *'-ę', '-e-sz', '-e', '-e-my', '-e-cie'* and *'-ą'* for the 1st, 2nd and 3rd person SG[1] and the 1st, 2nd and 3rd person PL, respectively. The themes which are members of conjugation II take the endings: *'-ę', '-i-sz', '-i', '-i-my', '-i-cie', '-ą'*. The verbs which belong to conjugation III have the endings: *'-m', '-sz', '-0', '-my', '-cie'* and *'-ą'*. Every conjugation is divided into classes, depending on the ending of the verbal theme. For instance, conjugation I consists of five classes and the fifth class consists of verbs which exhibit unproductive, idiosyncratic alternations.

Every Polish noun is inflected according to an extensive declensional paradigm. Each paradigm consists of seven cases. There are three declensional groups for SG nouns: masculine, feminine and neuter and two for PL nouns - nonpersonal and personal (Zagórska-Brooks, 1975).

Fully two-thirds of words in Polish exhibit changes which are consonantal and/or vocalic. Broadly speaking there are two sources of alternations in Polish: phonological and morphophonological (Zagórska-Brooks, 1975; Kowalik, 1998). Our focus here is on the second type, the morphophonological alternations. This type of alternation is caused by an interaction between morphology and phonology during concatenation of morphemes with other morphemes (e.g. in conjugation or declension). There are two kinds of morphophonological alternations in Polish verbs and nouns: regular and irregular. The regular alternation is predictable and usually productive. For instance, the postalveolar consonants - [ʃ] and [ʒ] in *pro**sz**-ę* [prɔʃ-ɛ̃] 'I ask' and *wo**ż**- ę* [vɔʒ-ɛ̃] 'I transport' alternate with the palatal consonants - *ś* [ɕ] and *ź* [ʑ] when the stem is concatenated with the vowel *-i-*, as in *pro**si**-ć* [prɔɕi-t͡ɕ] 'to ask' and *wo**zi**-ć* [vɔʑi-t͡ɕ] 'to transport'. The reason for this is that the alveolar place of articulation in the inflected forms (1st person SG, present tense) changes to the palatal place of articulation, because of the following front vowel *i*. In contrast, the irregular alternations cannot be predicted and are usually unproductive. For instance, the stem ***trze**-ć* [tʃɛ-t͡ɕ] 'to grate' alternates with the stem ***tr**-ą* [tr-ɔ̃] 'they grate'. There are two kinds of alternation here: the consonantal and the vocalic, where the vocalic alternation e : 0 is an idiosyncratic one, which is unpredictable from the morphophonology of contemporary Polish.

The fact that the morphophonological alternations in Polish verbs have a different distribution to those found in English is worth emphasising. In English, alternations in verbs occur between the present and past tense forms, for instance *give/gave*. In Polish, on the other hand, there can be several different alternations in verbs in the same and different tenses, depending on person, number and so forth. For the verb *wieś-ć* [vɛ́ɕ-t͡ɕ] 'to lead', there are two types of alteration: consonantal and vocalic. There are three consonantal alternants: *wie**ś**-ć* [vɛ́ɕ-t͡ɕ]: *wio**d**-ą* [vio**d**-ɔ̃]: *wie**dzi**e-sz* [vie**d͡ʑ**ɛ-ʃ] 'to lead: they lead: you lead'. And there are also three vocalic alternants, two in the present tense: wi**e**ś-ć: *wi**o**d-ą* and the third in the past tense: *wi**ó**d-ł* [vʲ**u**d-w] 'he led'.

[1] The following abbreviations are used: SG - Singular number, PL - Plural number, NOM - Nominative, DAT - Dative, LOC - Locative, VOC - Vocative;

The alternations in Polish have a more even distribution than in English. All classes, despite regular or irregular alternations, concatenate with the same conjugational endings, which are characteristic for a given conjugation. Polish is very similar to French in this respect. This does not occur in English, where the concatenative process applies to regular past tense forms but not to irregular ones. Neither does it happen in Italian, where the irregular forms conjugate using a different set of endings than the regular forms.

Experiment 1

The aim of Experiment 1 was to investigate processing of Polish verbs and nouns with regular and irregular alternations. We chose cross-modal priming as a starting point because most cross-linguistic and English data come from this paradigm. The question that we asked was whether the pattern of results for Polish items with no alternations, regular alternations, and irregular alternations would be comparable with the Italian and French or with the English and German results. In contrast to all of these languages, Polish also provides an opportunity to investigate the processing of regular and irregular nouns which are inflectionally suffixed; the suffix occurs in both singular and plural and denotes case, gender and number. Taking advantage of this characteristic of Polish nouns, we contrasted verbs and nouns. This served firstly, to test whether the representation of verb and noun alternants in Polish correspond to each other, and secondly, to obtain a broader picture of the representation of alternants cross-linguistically.

Irregularity in Polish verbs is defined on the basis of the alternation which occurs in the stem and not on the basis of which conjugational morphemes are selected for a given verb. The difference between regular and irregular verbs lies in that regular verbs exhibit an alternation in the stem which can be derived by the phonological rules of contemporary Polish or is conditioned by a specific morphophonological context. For irregular verbs, the alternations in the stems cannot be predicted. All Polish nouns belong to one of the declensional paradigms and their membership depends on the gender of a noun and on the last consonant of the theme. Irregularity and regularity in Polish nouns is defined in the same way as for Polish verbs[2].

To test the processing of words with alternation we covaried regular and irregular alternations in verbs and nouns and contrasted them with verbs and nouns without alternations. Conditions 1 to 3 were concerned with verbs, while conditions 4 to 6 were concerned with nouns. **Condition 1** consisted of verbs with no alternations, such as *czyt-a-sz/czyt-a-ć* 'you read/to read'. The aim here was to select as homogenous a group of items as possible, to serve as a baseline for verbs. **Condition 2** consisted of verbs with regular alternations, which can be predicted from the morphophonological rules of contemporary Polish, as in *nos**sz**-ę/nos**-i**-ć* 'I carry/to carry'. The alternation involved dental consonants, such as *c, dz*; postalveolar consonants, such as *sz, ż;* and palatal consonants - *ć, dź, ś, ź*. The reason for selecting this particular type of alternation is that it is regular and occurs in verbal stems which belong to a productive conjugation. This contrasts with the verbs in **Condition 3** with irregular alternations, as in *trz-e-ć/tr-ą* 'to grate/they grate'. Six types of irregular alternation were included in this condition[3]: **1)** a : e, **2)** ą : n' and ą : m', **3)** a : 0, **4)** o : e, **5)** ą : ę and **6)** e : 0. The motivation for selecting these types of alternation was that they could not be predicted from the morphophonological rules of Polish. These alternations are peculiar to a specific group of verbs, which belong to the unproductive classes of conjugation I.

Conditions 4 to 6 concentrated on nouns. **Condition 4** consisted of nouns with no alternation, e.g. *plac-u/plac* 'a square, GEN, LOC, VOC/a square, NOM, ACC' and served as a baseline for the other noun conditions. **Condition 5** was designed to investigate the representation of nouns with regular alternations. It consisted of nouns with three types of alternation, all of which are predictable from the context. These included the alternation of hard labials to soft labials, as in *chłop/chłop-i* 'peasant/peasants', the alternation of coronals, as in *studen**ci**-e/studen**t*** 'a student, LOC, VOC/a student, NOM', and the alternation of velars, as in *no**dz**-e /no**g**-a* 'a leg DAT, LOC/a leg, NOM'. **Condition 6**, in contrast to Condition 5, concentrated on alternations which are not fully predictable from the morphophonological rules of Polish. Three types of irregular alternation were included: **1)** a : e, **2)** ę : ą and **3)** o : ó. All three alternations exist nowadays as fossilised historical remnants. The alternations ę : ą and o : ó are based on a very old phonetic change.

Finally **Condition 7** consisted of words with phonological overlap, without semantic or morphological relationship, e.g. ***kot**let/**kot*** 'cutlet/cat'. This was designed to test whether priming could be due just to pure phonological overlap between the prime and the target.

Procedure

We used the cross-modal immediate lexical decision task. The participants heard binaurally an auditory prime, at the offset of which they immediately saw a visual target (for 500 ms) and had to decide, by pressing an appropriate button, whether a target word was a real word or a non-word. The subjects were allowed 2500 ms from the onset of the target for their response.

Targets were preceded either by a related prime or by an unrelated control prime. The control prime was matched for lemma frequency and number of syllables to the related prime. The priming effect was measured as a difference in the reaction time to the target word when preceded by the control prime as opposed to the related prime. Each target only occurred once in the experiment, and two experimental versions were run, alternating control and related primes for each target.

---

[2] There is also a second type of irregularity in Polish nouns which originates from the fact that nouns which are for instance masculine, e.g. *mężczyzn-a* 'a man' are inflected in SG according to one of the feminine declensional paradigms. This type of irregularity was not the focus here.

[3] The vocalic alternations were in some cases accompanied by consonantal alternations. For instance an alternation of *rz : r* co-occurs with *e : 0* in *trz-e-ć/tr-ą* 'to grate/they grate'. Pairs of this type were included here, because there are very few verbs with an irregular alternation which are not accompanied by the consonantal alternation.

The experiment was run in Kraków using DMASTR/VMASTR software[4]. A 486/33 NIMBUS PC was used for running the experiment and collecting participants' RTs and error rates. A Digital Audio Tape Corder TCD-D3, Sony Walkman was used to present the auditory primes.

Results

Three subjects from version 1 and one subject from version 2 were discarded from the analyses because of a high error rate in the lexical decision task. This gave 21 subjects per version. All participants were Polish native speakers who were studying in Kraków. The majority of them were in their twenties and some were in their thirties. Out of the total of 580 items per version, 5 experimental pairs had to be discarded from the analyses; two because of high error percentage, two because they were erroneously classified as having a certain type of alternation, and one item because of homophony. Every data point has been inversely transformed to reduce the influence of outliers. (see Table 1 for details of the descriptive statistics).

Table 1. Priming effects and error rates in Experiment 1.

Condition	N	Prime (Mean RT)	Control (Mean RT)	Priming
1 No Alternation Verb	20	527 (1.0)	610 (5.5)	83***
2 Regular Alternation Verb	26	534 (0.8)	622 (3.4)	88***
3 Irregular Alternation Verb	26	542 (1.1)	615 (4.2)	73***
4 No Alternation Noun	20	512 (0.5)	573 (3.3)	61***
5 Regular Alternation Noun	30	518 (0.5)	580 (3.5)	62***
6 Irregular Alternation Noun	30	503 (0.7)	549 (1.5)	46***
7 Phonological Overlap	20	581 (6.5)	571 (3.5)	-10

*** denotes p<0.01; Reaction Times are in ms; Error rates in percentages (in parentheses).

An overall ANOVA with Prime (2 levels), Condition (7 levels) and Version (2 levels) was run, separately for subjects (F1) and items (F2), and revealed that there was a facilitatory effect of Prime F1(1, 40)=160.22, p<0.001, F2(1, 153)=200.94, p<0.001. There was a main effect of Condition F1(6, 240)=35.52, p<0.001, F2(6, 153)=3.94, p<0.01, and a two-way interaction of Condition and Prime F1(6, 240)=160.22, p<0.001, F2(6, 153)=8.67, p<0.001.

Subsequently an analysis of the simple effects of Prime at each level of the Condition was run. There was a consistent facilitatory effect of Prime at all levels of Condition with the exception of the Phonological Overlap Condition (p<0.001 throughout). To explore whether there were any differences in the magnitude of priming between no alternation, regular alternation and irregular alternation, planned comparisons were run separately for verbs and nouns. Regarding verbs, there was no significant difference in the magnitude of priming between No Alternation and Regular Alternation (F1<1, F2<1), No Alternation and Irregular Alternation, (F1<1, F2<1), and Regular Alternation and Irregular Alternation, F1(1, 40)=2.866, p=0.098, F2(1, 46)=1.31, p=0.258. The results for the nouns paralleled those of the verbs. There was no significant difference between the magnitude of priming for No Alternation and Regular Alternation, (F1<1, F2<1), No Alternation and Irregular Alternation, [F1(1,40)=1.281, p=0.264, F2<1] and Regular Alternation and Irregular Alternation, F1(1, 40)=1.208, p=0.278, F2<1.

Discussion

Significant priming was found for regulars as well as for irregulars, and no differences were found in the magnitude of priming between all verb conditions compared together, and all noun conditions compared together. Although there was a slight numerical decline in the amount of priming for the irregular alternations, this was not statistically significant in any analyses.

These results suggest, first, that the representation and processing of verb and noun forms is comparable in Polish. The only difference here is the significantly larger amount of priming for verbs overall compared with nouns. A possible explanation for this is that nouns in the oblique cases (all noun primes had this characteristic) are less effective primes than conjugated verbs (all the verb primes were conjugated forms). Nouns in the oblique case typically occur in a specific prepositional context, and they may be more difficult to process out of context, in contrast to conjugated verb forms, which can frequently occur without further context.

More generally, there is again no evidence here that the processing of Polish regular and irregular verbs and nouns invokes distinct underlying processing. This finding groups Polish data together with the data on Italian and French, but not with the English and German results. However, before proceeding further, we need to put these results on firmer ground.

The absence of priming for the pairs with pure phonological overlap indicated that the effects cannot be attributed just to phonological overlap. We cannot, however, exclude an account purely in semantic terms, since all of the verb and noun test pairs were strongly related not only morphologically but also semantically. It is possible, for example, that the preserved priming for the irregular alternants reflects a strong semantic input, compensating for reduced morphological priming. To address this concern, we ran a subset of the items from Experiment 1 (the verb materials) in a second experiment designed to separate morphological and semantic effects.

Experiment 2

This experiment uses the delayed repetition auditory-auditory priming task. In this task, where several items intervene between prime and target, we generally see no effects of pure semantic relatedness but robust effects of morphological relatedness (e.g., Marslen-Wilson & Tyler,

[4] DMASTR/VMASTR software was developed by Ken and Jonathan Forster at the University of Arizona, Tucson, U. S. A.

1998). If the priming effects for the alternation conditions are primarily morphological in character, then these should survive the change in task.

Materials

Because of the requirements of the task, the inclusion of all the materials from Experiment 1 would have made the experiment infeasibly large. We therefore used only the verb materials, which can be linked most directly to the research in other languages.

108 prime-target pairs were selected for 5 experimental conditions. Three of these were the Verb conditions from Experiment 1, with 20 pairs in the No Alternation condition, 24 in the Regular Alternation condition, and 24 in the Irregular Alternation condition. A further 20 pairs were selected from the Phonological Overlap condition (***kot***/*let*/***kot*** 'cutlet/cat'). The fifth, new condition consisted of 20 pairs of words that were Semantically but not Morphologically Related [+Sem, -Morph] (*banan/kokos*, 'banana/coconut'). This was to check whether purely semantic priming could be observed.

Procedure

Standard delayed auditory-auditory priming was used. The subjects heard binaurally a string of words, one every 3 seconds, which were a mixture of primes, controls, targets and fillers. Their task was to make a lexical decision by pressing an appropriate button on a response box to every heard item. The task was designed in such a way that primes and targets were separated by 12 intervening items (approximately 35 seconds). The experiment was run in Kraków using DMASTR/VMASTR software. The same experimental equipment as in Experiment 1 was used.

Results

4 participants from version 1 and 3 from version 2 were discarded from the analysis on the same criteria as in Experiment 1. Additionally 1 participant from version 1 and 1 from version 2 were excluded, because of technical problems. A total of 19 subjects per version was entered into the analysis. Further details on participants were the same as in Experiment 1. Two experimental items were removed from the analysis, because of high error percentages. A total of 106 items were entered into the analysis. (See Table 2 for the details of the descriptive statistics).

The analyses are based on the Target reaction times only. The inversely transformed data were analysed in a repeated measures ANOVA. The overall analysis for all 5 conditions revealed that there was a main effect of Prime, $F1(1, 36)= 40.14$, $p<0.001$, $F2(1, 96)=16.17$, $p<0.001$. There was also a main effect of Condition, $F1(4, 144)=67.80$, $p<0.001$, $F2(4, 96)=4.24$, $p<0.01$ and a two way interaction of Condition and Prime $F1(4, 144)=7.26$, $p<0.001$, $F2(4, 96)=3.50$, $p<0.05$.

Subsequently, the simple effects of Prime on each level of the Condition were investigated. There was significant priming in all the verb conditions as predicted: No Alternation Verb: $F1(1, 36)=11.05$, $p<0.01$, $F2(1, 16)=5.69$, $p<0.05$; Regular Alternation Verb: $F1(1, 36)=25.97$, $p<0.001$, $F2(1, 22)=12.69$, $p<0.01$; Irregular Alternation Verb: $F1(1, 36)=28.99$, $p<0.001$, $F2(1, 22)=9.56$, $p<0.01$. No significant priming was found either in the Phonological Overlap condition [$F1(1, 36)=1.41$, $p=0.242$, $F2<1$], or in the Semantically Related and Morphologically Unrelated condition [$F1(1, 36)=1.83$, $p=0.184$, $F2<1$].

To explore whether there were differences in the magnitude of priming between the three verb conditions the appropriate planned comparisons were made. There were no significant differences in the magnitude of priming in any of the comparisons between the No Alternation, Regular Alternation, or Irregular Alternation conditions ($F1<1$, $F2<1$ throughout).

Table 2. Priming effects and error rates in Experiment 2.

Condition	N	Prime (Mean RT)	Control (Mean RT)	Priming
1 No Alternation Verb	20	718 (0.6)	748 (0.9)	30*
2 Regular Alternation Verb	24	743 (1.1)	781 (1.8)	39**
3 Irregular Alternation Verb	24	753 (3.7)	785 (3.7)	32**
4 Phonological Overlap	20	816 (8.9)	803 (7.4)	-14
5 [+Sem, -Morph]	20	810 (2.6)	819 (1.8)	9

* denotes $p<0.05$; ** denotes $p<0.01$; Reaction Times are in ms; Error rates in percentages (in parentheses).

Discussion

The results indicate that the priming effect observed in all the verb conditions is morphological in nature and not semantic. As in Experiment 1, the effects are equally strong across the alternation sets. There is no trace of a semantic effect under these testing conditions, but there is robust priming between all morphologically related pairs, indicating that both prime and target map onto the same underlying morpheme at the level of the lexical entry. The absence of significant priming for the Phonological Overlap condition confirms that any facilitatory priming here cannot be attributed to the pure phonological overlap between primes and targets.

General Discussion

The question asked at the beginning of this paper was whether a dual or single mechanism can best account for the results from Polish. The data we report, with no difference in the magnitude of priming for regulars and irregulars, seems to group with earlier findings on Italian and French (Meunier & Marslen-Wilson, 2000; Orsolini & Marslen-Wilson, 1997), and indicates that a uniform set of processing procedures and representations are applied to Polish regular and irregular verbs and nouns.

We suggest that the explanation for these properties of Polish, and its similarities with French and Italian, lies in the fact that every verb and noun in Polish functions within similar inflectional paradigms which requires suffixation. All verbs and nouns, despite having regular or irregular vocalic or consonantal alternations, are inflected using con-

catenative processes which link verb roots to a series of inflectional suffixes. This is in contrast to English, where the regularity/irregularity distinction is confounded with a contrast in type of morphological process. The regular past tense inflection involves concatenative suffixing morphology, whereas the irregular forms have no overt inflectional structure, and have to be both stored and processed as unanalysable full forms. The sequence *jumped* can be broken down into the pair {jump} + {-ed}, but no such analysis can be applied to forms like *bought* or *gave*.

The consequences of this, in English, is that regular and irregular inflected forms do partially involve separable underlying cognitive and neural systems, with the regular inflected forms requiring the involvement of specialised systems supporting processes of morphophonological assembly and disassembly. These systems, that can be selectively damaged following stroke and other injuries to the brain, are not required for the access and processing of the irregular forms, which are not phonologically decomposable (Marslen-Wilson & Tyler, 1998). The same distinction does not apply, of course, to Polish (or French and Italian) irregulars, all of which are morphophonologically complex, and which invoke the same range of processing mechanisms as the regular forms.

These considerations suggest that further progress in elucidating the properties of the mechanism underlying the processing of regular and irregular forms will require more attention to the neural underpinnings of these mechanisms. A start has been made in this direction in the study of English and German inflection. For instance, Marslen-Wilson, Csibra, Ford, Hatzakis, Gaskell and Johnson (2000) examined the processing of English regular and irregular past tenses using an immediate cross-modal priming experiment during which ERPs were recorded. The results revealed differences in the pattern of scalp activity for regulars and irregulars and provide support for the claim that partially different mechanisms underly the processing of each type of material. Two other ERP studies reported by Clahsen (1999) on the representation and processing of German regular and irregular plurals and participles, using a violation paradigm, claim that regularisations elicit signals which are characteristic for violations of a morpho-syntactic rule, whereas irregularisations elicited signals which are typical of anomalous words and can be regarded as support for a distinction between rule-based and lexically based inflectional processes (Pinker, 1999).

Thus, if Polish regular and irregular nouns and verbs are indeed processed by a uniform underlying mechanism, as suggested by the data presented in this paper, then one would not predict significant differences in ERP recordings for the regulars and irregulars, nor in neuro-imaging studies using techniques such as fMRI, allowing a greater degree of spatial localisation. But whatever the outcome, input from studies of this sort are likely to be an essential component of any future resolution of the kinds of questions raised here.

Acknowledgements

This research is supported in part by an ESRC research studentship to A. Reid, and in part by the UK MRC.

References

Clahsen, H. (1999). Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral & Brain Sciences*, 22, 991-1060.

Kowalik, K. (1998). Morfonologia. In R. Grzegorczykowa, R. Laskowski & H. Wróbel (Eds.), *Gramatyka współczesnego języka polskiego. Morfologia.* Warszawa: PWN.

Laskowski, R. (1998). Paradygmatyka. Czasownik. In R. Grzegorczykowa, R. Laskowski & H. Wróbel. (Eds.), *Gramatyka współczesnego języka polskiego. Morfologia.* Warszawa: PWN.

Marslen-Wilson, W. D., Csibra, G., Ford M., Hatzakis, H., Gaskell G. & Johnson, M. (2000). Associations and dissociations in the processing of regular and irregular verbs: ERP evidence. Poster presented at *The conference of Cognitive Neuroscience Society*, San Francisco.

Marslen-Wilson, W. D. & Tyler, L. K. (1998). Rules, representations, and the English past tense. *Trends in Cognitive Science*, 2 (11), 428-436.

Marslen-Wilson, W. D, Hare, M. & Older L. (1995). *Priming and blocking in English inflectional morphology*. Experimental Psychology Society, London, January 1995.

Meunier, F. & Marslen-Wilson, W. D. (2000). Regularity & irregularity in French inflectional morphology. *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society* (pp. 346-351). The University of Pennsylvania in Philadelphia, PA., Mahwah, N.J: LEA.

Orsolini, M. & Marslen-Wilson, W. D. (1997). Universals in morphological representation: Evidence from Italian. *Language & Cognitive Processes*, 12(1), 1-47.

Pinker, S. (1999). *Words and Rules*. London: Weidenfeld & Nicolson.

Pinker, S. & Price, A. (1991). Regular and irregular morphology and the psychological status of rules of grammar. *Proceedings of the 17th Annual Meeting of the Berkeley Linguistic Society* (pp. 230-251).

Plunkett, K. & Marchman, V. (1993). From rote learning to system building: acquiring verb morphology in children and connectionist nets. *Cognition*, 48, 21-69.

Rumelhart, D. E. & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J.L. McClelland, D.E. Rumelhart, & PDP Research Group, *Parallel distributed processing: explorations in the microstructure of cognition.* Volume 1. Cambridge, MA: Bradford Books/MIT Press.

Sonnenstuhl, I., Eisenbeiss, S. & Clahsen, H. (1999). Morphological priming in the German mental lexicon. *Cognition*, 72, 203-236.

Stanners, R. F., Neiser, J. J., Hernon, W. P. & Hall, R. (1979). Memory representation for morphologically related words. *Journal of Verbal Learning & Verbal Behaviour*, 18, 399-412.

Zagórska-Brooks, M. (1975). *Reference grammar*. The Hague: Mouton & Co. N.V. Publishers.

Cats could be dogs, but dogs could not be cats: what if they bark and mew? A Connectionist Account of Early Infant Memory and Categorization

Robert A.P. Reuter (rreuter@ulb.ac.be)
Cognitive Science Research Unit, CP 191, Av. F.D. Roosevelt, 50
B-1050 Brussels, Belgium

Abstract

The goal of this paper is to replicate and extend the connectionist model presented by Mareschal and French (1997) as an account of 'the particularities of […] infant memory and categorization'. With infants, the sequential presentation of cats followed by dogs yields an expected increase in infants' looking time, whereas the reversed presentation order does not. This intriguing asymmetry of infants' category formation, first reported by Quinn, Eimas, and Rosenkrantz (1993), was simulated by Mareschal et al.'s simple connectionist network. In addition, the authors proposed that this asymmetric categorization is a natural byproduct of the 'asymmetric overlaps of the visual feature distributions' of cats and dogs. Using a simple feedforward backpropagation network, we successfully replicated this asymmetric categorization effect, as well as a reported asymmetric exclusivity effect in the two categories, and an asymmetric interference effect of learning dogs on the memory for cats, but not of learning cats on the memory for dogs. We furthermore investigated the authors' explanation of the asymmetric effects, firstly, by systematically varying the overall similarity between learned items and interfering items, and secondly, by adding a binary feature to the input set, namely the animal cry (barking vs. mewing). The results of the present modeling underscore the authors' explanation of the observed effects in infants' memory and categorization, but also suggest lines of further experimental research susceptible to undermine the proposed connectionist account.

Introduction

In this paper, we report on a replication and two extensions of Mareschal and French's (1997) simple connectionist model that accounts fairly well for unexpected findings observed in spontaneous category formation in young infants (e.g., Quinn, Eimas, & Rosenkrantz, 1993) and in infant memory. Indeed, Mareschal et al. focused their modeling efforts on three target behaviors of very young infants concerning their categorization and memory, namely: (a) the ability to categorize complex visual stimuli, (b) the asymmetric effect in early categorization, and (c) interference effects in early memory.

The empirical data Mareschal et al. modeled show that infants, aged 3 to 4 months, are able to accurately categorize cats and dogs, and that they form an exclusive category of cats (thus excluding novel dogs) but an inclusive category of dogs (thus novel cats may well fall into the category of dogs) (Quinn et al., 1993). Furthermore, infants are known to present catastrophic forgetting of previously learned stimuli when shown certain other intervening material (e.g., Cohen & Gelber, 1975). This interference effect decreases precisely when infants' categorization abilities increase (Quinn & Eimas, 1996). Infant memory and categorization can thus be thought to be closely linked to each other and to depend on the same basic mechanisms.

Indeed, the connectionist account given for the observed asymmetric effects in categorization and for the interference effects in memory is based on the same reasoning. The asymmetry in category formation arises from the unequal overlap of the visual features distribution and not just the variance of the distribution itself. In other words, the values of the cat features fall within those of the dog values. Hence, based merely on the statistical structure of the input features, infants (and neural networks) form a category of dogs including cats, whereas they form a category of cats excluding (some) dogs[1]. The correlational structure extracting mechanism is precisely what accounts for the observed asymmetric effects of unequally exclusive (or inclusive) cat and dog categorizations. Catastrophic forgetting, on the other hand, can be understood as the deleterious effect of representing, within the same connections, stimuli whose features are very differently distributed. This, consequently, "washes out" the relevant knowledge previously stored in the network. However learning items whose features lie within the same range as those learned previously should not create very different internal representations - and may hence even consist in "learning more of the same". Based on this connectionist account of the relationship between infant categorization and memory, Mareschal et al. successfully produced the conjectured asymmetry in catastrophic interference consistent with the asymmetry observed in category elaboration (i.e., dogs interfered with previously learned cats, whereas cats did not interfere with learned dogs).

[1] Indeed, some dogs "look" like cats, since that they fall into the range of the cat distribution for some features. We will come back to this later.

In the following we, first of all, report on a conceptual replication of Mareschal et al.'s modeling results, concerning the development of categories, the asymmetry in category exclusivity, and the asymmetry in interference effects. Next, we show that the results of a cluster analysis performed on the input data and on the hidden units activation patterns (after learning of all items had occurred) suggest that the asymmetry in category exclusivity closely depends on the particular items used for "cross-category compatibility"[2] testing. Then, we explain how we tested Mareschal et al.'s connectionist account of asymmetry in category exclusivity and in interference effects, based on two systematic manipulations of the overlap in feature distributions of cat and dog categories. The first variation of overlap was produced by carefully choosing the items presented for training and those for interfering. The second variation of overlap was produced by the introduction of a supplementary (binary) input feature that unambiguously, taken individually, separates cats from dogs. Finally, we propose lines of further experimental research that might undermine the connectionist account embraced here.

The Model

We made the same assumptions as Mareschal et al. about the mapping between experimental results found in infants with the technique of preferential looking times and modeling results. The increase of error in the model's output is indeed assumed to be related to increased infant attention. The results reported below are based on the performance of a standard 10-8-10 feedforward backpropagation network, as well as that of a standard 11-8-11 network when the supplementary input feature was added. For both models, learning rate and momentum were both set to 0.9. The Fahlmann offset used in the original paper was not used in our model. Networks learned the data for a maximum of 250 epochs or until all output bits were within 0.2 of their targets. Results are averaged over 50 replications. Weights were updated after each stimulus presentation. Further details about changes in procedure compared to Mareschal et al.'s model are given in the results section below.

The Data

The data were identical to those used by Mareschal et al. A description of their origin and their characteristics can be found in their paper. In the following paragraphs, we report a cluster Analyses performed on the input data (and on the activation patterns of the trained network's hidden units) that will provide some insight into the inherent structure of the input data and its implication on modeling results. They will as well allow us to justify the proposed extensions of Mareschal et al.'s neural network simulations.

Cluster Analysis

On the basis of the clusters of correlated values of the input data, cats and dogs cannot be clearly separated into two mutually exclusive categories (see figure 1). Likewise, the connectionist model's internal representations (as reflected by the activation pattern of its hidden units after training on the whole set of stimuli) do not reflect clear-cut groupings of just dogs or just cats (see figure 2). Indeed, some cats, respectively some dogs, are more prototypical in terms of their overall feature similarity with the other members of their category.

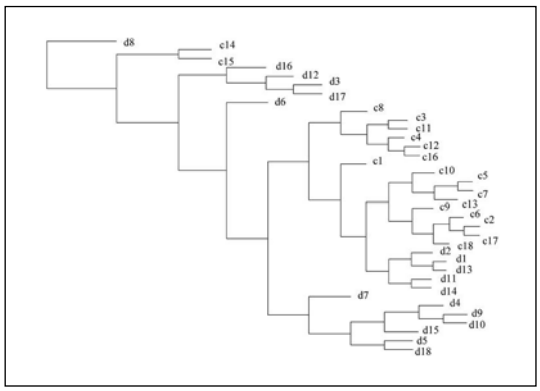


Figure 1. Cluster analysis performed on the input data patterns for cats (c1-c18) and dogs (d1-d18). Distances and cluster structure in this graph correspond to the overall similarity structure of the input patterns.

Moreover, and this is most crucial for Mareschal et al.'s proposed account of infant memory and categorization, nearly all cats (except 2 out of 18) fall into the "family" of "dog-cat"s[3], whereas, only 5 dogs (out of 18) clearly fall into the group which the majority of cats belong to. The clustering of cats and dogs into different subgroups suggests that the observed category exclusivity effects, as well as interference effects, should not occur for all combinations of learned and novel items (respectively, learned, interfering and novel/same-category items). Based on these results we predicted, precisely, that the closer items are in terms of cluster analysis distance, the smaller interference and cross-category exclusion effects should be.

[2] Cross-category compatibility refers to the question whether a novel exemplar is "accepted" as a member of the opposite category or not.

[3] "Dog-cat"s corresponds to a regrouping of clusters containing cats and dogs.

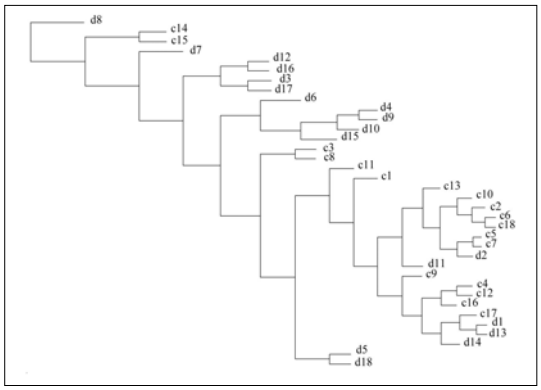


Figure 2. Cluster analysis performed on the activation patterns of the network's hidden units for cats (c1-c18) and dogs (d1-d18). Distances and cluster structure in this graph correspond to the overall similarity structure of the network's internal representations of the input patterns.

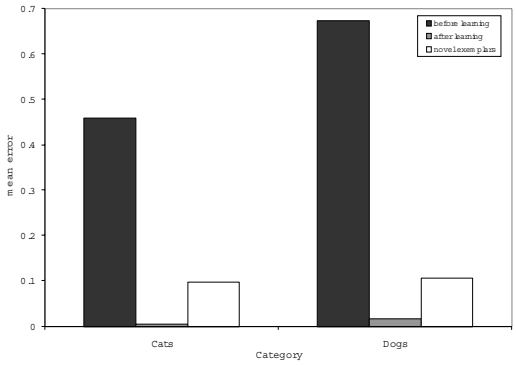


Figure 3. Mean error on the network's output when (a) presented with exemplars before learning, (b) presented with trained exemplars after learning, and (c) presented with untrained exemplars after learning.

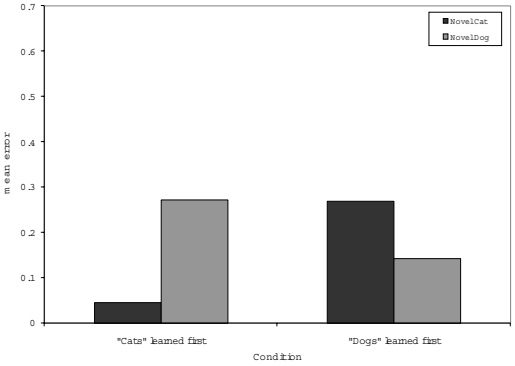


Figure 4. Asymmetric exclusivity of the cats and dogs categories. When trained first on cats, an untrained dog results in a larger increase of error than an untrained cat, but when trained on dogs, an untrained cat only produces a small increase in error as compared to an untrained dog.

Results

In the following section, we will briefly report the basic replication results since they nearly reflect the findings of Mareschal et al., and then describe our new results with their model.

The Development of Cat and Dog Categories

We obtained results comparable to those described in Mareschal et al. Networks do form a category of both cats and dogs. Figure 3 shows the initial mean error score, the mean error after training on the first 12 items of each category[4], and the mean error score (after learning) for the 6 remaining exemplars of the corresponding category (same-category testing). Needless to say that networks develop a faithful internal representation of both cats and dogs, and that they nevertheless recognize novel items as unfamiliar (small increase in error compared to the learned items). It is worth noting however that the initial error scores are slightly different between dogs and cats. Relative to those of cats, the features of dogs are more variable. Thus the mean error on dogs without training tends to be bigger than on cats. This confirms Mareschal et al.'s data analysis in terms of means and variances of the feature distributions of cats and dogs.

The Exclusivity of Cat and Dog Categories

Figure 4 shows the mean error on output of networks trained on 17 (out of 18) cats when they are presented with either the single remaining cat or with any of the dogs (18 out of 18), as well as with the corresponding opposite configuration (dogs learned first and then tested on novel dogs, respectively novel cats).

We used this training regime, instead of training the networks on 12 out of 18 items and then testing it on 4 cats, in order to get a reasonable number of different controlled combinations of training and testing items. All cats, respectively dogs, appear once as the novel item which the network has to categorize based on its nearly perfect "knowledge" of the same or opposite category. Our results show that, on average, dogs are less likely to be accepted as members of the cat category, than vice-versa. Indeed, a novel cat presented to a network that has learned dogs is less likely to produce a big increase in error, than a dog when presented to a network that was trained on cats. Based on the results of the cluster analysis, we suggest that the exclusion of the various cross-category items (but also of very atypical same-category items) somehow depends on their similarity with the core representation the network has developed during training.

[4] We suppose that this selection is pseudo-random, since we did not give the stimuli set a particular a priori order.

The Asymmetric Interference Effect

In this section we examine the effect of learning items of a second category on the networks ability to correctly "accept" novel items as belonging to the category it was trained on in the first time. The network, for instance, was trained on 12 cats, then tested on the remaining (novel) cats, then trained on 4 dogs, and finally re-tested on the novel cats. The difference in error scores on the novel cats before and after learning the 4 dogs is called interference effect, interference of dogs on the memory for cats. The same reasoning holds for the opposite case, memory for dogs interfered by training on cats. The interference effect of dogs on cats was easily replicated, by pseudo-randomly (see above) choosing cats to learn and dogs to interfere with the memory for cats. However, in contrast to Mareschal et al., we did observe a considerable interference effect of learning cats upon the memory for dogs.

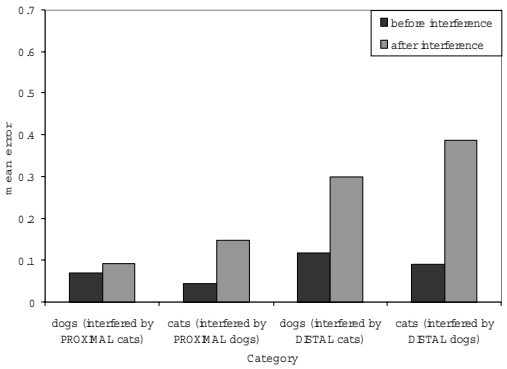


Figure 5. Network performance with untrained exemplars before and after learning an interfering category as a function of overall similarity between trained items and interfering items (distal vs. proximal).

See figure 5 for the results on the network's performance with novel items before and after learning items of the opposite category. We thus observed catastrophic interference in both cases, when not controlling for the similarity between training, testing and interfering. In figure 5, "distal" refers to items that are very dissimilar to the target category based on the cluster analysis. We thus analyzed the similarity between training and interfering items, we had pseudo-randomly chosen, and furthermore carefully selected packages of items in order to produce the "desired" results, based on the distance between items estimated by cluster analysis. The results of the simulations conducted on these items is reported in the following paragraph.

The Effect of Similarity between Learned and Interfering Items

We trained the network with various combinations of category learning items and interfering items from the other category. The choices of the items were motivated by the results of the cluster analysis performed on the input data. This enabled us to distinguish between groups of stimuli that are more or less "similar" to each other. We predicted that interference depends upon the overall feature similarity (i.e., overlap of feature distributions) between the central tendencies of these two groups, the bigger the similarity, the smaller the interference should be and vice-versa. We were, by this subterfuge, able to qualitatively reproduce Mareschal et al.'s asymmetry in interference effects by showing that, under selected conditions, dogs are not interfered by cats (see figure 5, dogs interfered by proximal cats). We were also able to show that results contrary to those reported in Mareschal et al. could be found. If cats learned in the first place, were "interfered" by very similar dogs, and the test items (i.e., novel cats) were chosen as close to the (two) learned set(s), then interference was minimal and relatively close to that observed in the case of dogs "not-interfered" by (similar) cats. Our results show that the results found by Mareschal et al. need not be the only possible ones. Furthermore, we could not manage to replicate the reported "average"[5] absence of interference of learning cats on the memory for dogs. Still, consistent with their connectionist account of infant category and memory our results are conclusive regarding the influence of overall feature similarity on the exclusivity and interference effect, though locally and not really globally. We could indeed show that overall similarity of items used for training and those used for interference could be used to predict interference of learning a second category on the memory for a first category. What we could not show was that this really is true for cats and dogs, on average, taken as categories. We remain agnostic to the very reasons of our failure to reproduce the absence of interference of cats on the memory for dogs when exemplars of both categories were randomly chosen. If the account given by Mareschal et al. is as general as they presume, a claim to which we subscribe in principle, then it should show up more consistently and more reliably with nearly any random selection of cats and dogs. Although we stay puzzled concerning the precise reasons of our failure to replicate, we have some hints on potential explanations. In fact, we wonder whether the rather small increase in error on dogs after interference by cats is consistent with the nevertheless not so small increase in error found with novel cats when dogs have been learned first. If it is true that, on average, learning cats does not have a deleterious influence on the memory for dogs, then why does presenting cat after training on dogs produce an increase in error compared to

[5] Since the authors did not precisely mention how they chose their items for training and interfering, we assume that the reported asymmetry in interference effects must be supposed to reflect average results, which is quite consistent with their account of the category exclusivity effects.

presenting a dog? After all, if learning cats is truly somewhat equivalent to learning more of the dog category[6], then why does testing on novel cats not produce less error then testing on novel dogs? We think that this prediction might be consistent with the connectionist account proposed by Mareschal et al., solely based on the asymmetry of overlap of feature distributions. Since nearly all cats belong to a narrow range of distribution embraced by the distribution of the dog category, dogs are more likely, on the average, to fall outside this narrow range, within which the "prototypical" dog also falls and which is precisely what the network's internal representation should reflect. Comparing cats and dogs to this "prototypical" dog should show that cats are situated closer to it in nearly all considered feature distributions, thus producing less error than dogs (which are more likely than cats to be different from this prototype).

The Effect of Induced Changes in the Inherent Structure of the Stimuli

We conducted the same simulations as before, except that networks were trained on inputs that included the animal cry (barking vs. mewing) as an eleventh characteristic variable. A cluster analysis (shown in figure 6) performed on these input data shows that this manipulation manifestly segregates dogs and cats into two next-to-perfectly distinct categories. The increase in distinctiveness of cats and dogs produced by the addition of the binary variable should eliminate (or at least considerably reduce), we hypothesized, the asymmetry of the category exclusivity, as well as the asymmetry of cross-category interference effects.

The results shown in figures 7 and 8 confirm our predictions based on the account given by Mareschal et al. In other words, cats no longer could be part of the dog category, and learning cats considerably interfered with the memory for the dog category. The inherent correlational structure of the input data thus has an important effect on networks categorization and memory. Note that the networks successfully formed a category of cats and dogs just like when the animal cry was not added to the stimuli features.

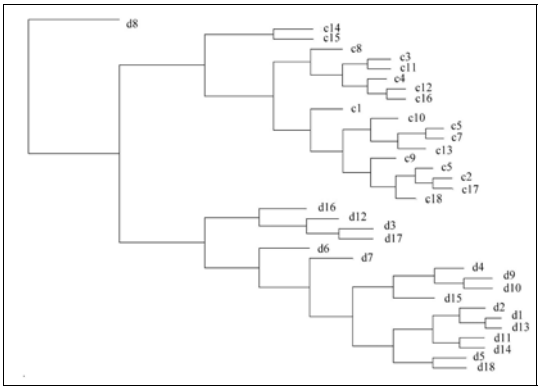


Figure 6. cluster analysis performed on the input data patterns for cats (c1-c18) and dogs (d1-d18) after addition of the "animal cry" feature. Distances and cluster structure in this graph correspond to the overall similarity structure of the modified input patterns.

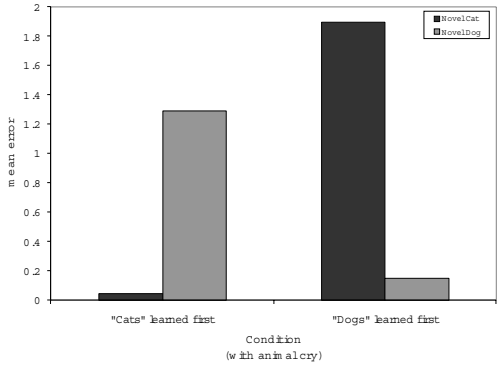


Figure 7. Symmetric exclusivity of the cats and dogs categories, when "animal cry" is added to the input features.

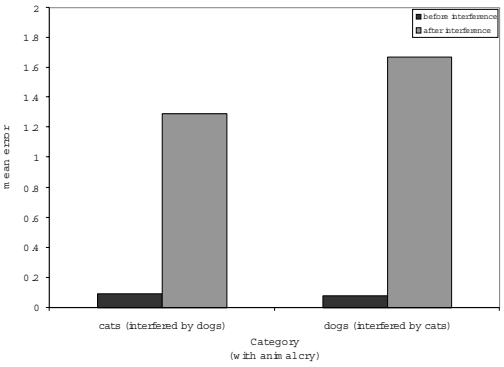


Figure 8. Network performance with untrained exemplars before and after learning an interfering category, when "animal cry" is added to the stimulus features.

[6] This is of course only true in some sense, since the dog category is clearly characterized by greater feature variance.

Discussion

Connectionist autoassociator networks, like young infants, form categorical representations of cats and dogs. The categories developed show asymmetric exclusivity, closely related to the unequal distribution of features in the stimuli shown. Most of the cats could be classified as dogs, but most dogs are not plausible cats. The model also suggests the presence of asymmetric interference effects of sequential learning of cats and dogs. Such effects have recently (Mareschal, French, & Quinn, draft) been observed in infants. Supported by empirical works on infants, the model thus pleads for a close link between the mechanisms underlying infant visual memory and categorization. In fact, Mareschal et al. (1997) claim that some kind of associative, data driven mechanism underlies early visual memory and categorization. The present paper underscores this claim, by showing that explicit manipulations of the correlational structure of the data input influences the networks performance. Connectionist models, by making clear assumptions about the input data, can thus be helpful in predicting which stimulus features are likely to be taken into account by infants. Indeed, experimental works (Spencer, Quinn, Johnson, & Karmiloff-Smith, 1997) have shown that infants typically rely 'on head/face information to categorically differentiate between cats and dogs, under conditions of short exposure duration'. This empirical result is consistent with the connectionist model presented here, since face visual features (viz. nose length and nose width) are the most informative features about the cat/dog distinction. In addition, based on Mareschal et al.'s (1997) connectionist account, it seems reasonable to predict that presenting pictures of cats and dogs in association with the corresponding animal cry should produce the same results in infants than in networks, namely mutual and symmetric category exclusivity and symmetric interference effects. Based on simulation results, we also predict a close parallelism between infants' and connectionist models' performances in memory and categorization task in terms of the similarity of particular stimuli (and combinations of stimuli). Truly, the model incorrectly excludes certain stimuli and not others, thus infants should present the same behavior pattern with precisely those stimuli in question. Likewise, if the model, for certain items but not others, does not present the discussed asymmetry in interference effects then infant's behavior should qualitatively reflect the same catastrophic forgetting. We thus suggest an item based analysis of networks' and infants' memory and categorization performances. Finally, we would like to recall that, by construction, connections networks' performance depends upon the very selection of certain stimulus features and not other. Thus, if networks produce categorization and memory effects similar to those of infants, then the selection of the particular features is given support. Nevertheless, it must be experimentally shown that infants actually rely on those features and not others. Still, connectionist models provide good predictions about which stimulus features are most likely to participate in infant categorization and memory.

Acknowledgments

Robert A.P. Reuter is a Research Assistant of the National Fund for Scientific Research (Belgium). Thanks to Bob French and Axel Cleeremans for their insightful comments on earlier drafts of this paper.

References

Cohen, L. & Gelber, E. R. (1975). Infant visual memory. In L. Cohen & Salapatek (Eds.), *Infant perception: From sensation to cognition*, Vol. 1 (pp. 347-403). NY: Academic Press.

Mareschal, D., & French, R. M. (1997). A Connectionist Account of Interference Effects in Early Infant Memory and Categorization, In *Proceedings of the 19^{ih} Annual Cognitive Science Society Conference*, NJ: LEA, pp. 484-489).

Mareschal, D., French, R. M., & Quinn, P. (draft, April 14, 1998). Interference Effects in Early Infant Visual Memory and Categorisation: A Connectionist Model.

Quinn, P. C., & Eimas, P. D. (1996). Perceptual organization and categorization in young infants. In C. Rovee-Collier & L. P. Lippsitt (Eds.), *Advances in infancy research* (Vol. 10, pp. 1-36). Norwood, NJ: Ablex.

Quinn, P. C., & Eimas, P. D., & Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants, *Perception*, 22, 463-475.

Spencer, J., Quinn, P. C., Johnson, M. H., & Karmiloff-Smith, A. (1997), Heads you win, tails you loose: Evidence for young infants categorizing mammals by head and facial attributes, *Early Development and Parenting*, Vol. 6 (3-4), 113-126.

Motor Representations In Memory And Mental Models: Embodiment in Cognition

Daniel C. Richardson (dcr18@cornell.edu)
Michael J. Spivey (mjs41@cornell.edu)
Jamie Cheung (jmc56@cornell.edu)
Cornell University, Department of Psychology
Uris Hall, Ithaca, NY 14853

Abstract

A variety of experimental results have suggested that motor systems can participate in what were thought to be purely perceptual tasks. Extending previous work in the stimulus-response compatibility paradigm, we show that a representation of a visual stimulus accessed from memory can activate potential motor interactions. Other research has shown that mental images and mental models can have analogue or spatial characteristics. In our second experiment, we present evidence that such representations, generated purely from linguistic descriptions, can also activate motor affordances. Results are discussed within the context of the embodiment of *conceptual*, as well as perceptual, processes

Introduction

One can often observe people who, having lost the set of keys they held a moment ago, mentally retrace their steps, miming their previous interactions with objects and places that might now be a hiding place. Similarly, when imagining the rearrangement of a room, people often move their hands as if picking up and moving furniture. The gestures that embellish discourse are also made by people who are on the telephone, and even by the congenitally blind (Iverson & Goldin-Meadow, 1998).

Hand waving examples such as these offer glimpses of the relationship between 'higher' cognitive functions such as language and imagination, and the seemingly more mundane perceptual and motor systems that carry out our daily chores. Increasingly, it is suggested that these motor systems have an important contribution to cognition.

This zeitgeist of 'embodiment' has been described as 'a seismic event … taking place in cognitive science' (Newton, 1998) and a 'perennial recycling of behaviourist ideology' (Pylyshyn, 2000). In general, proponents of an embodied perspective reject the idea of cognition as wholly the processing of abstract, or amodal, symbolic representations. They emphasise the ways in which the representational and processing burden of cognition can be offloaded on the external world and the motor and perceptual systems that interact with it (Barsalou, 1999a; Clark, 1997; Lakoff 1999)

There is a growing weight of behavioural and brain imaging work that implicates motor systems in perceptual judgements. Observers often interpret visual stimuli in terms of the physically plausible motions it would take to produce them (eg Shiffar & Freyd, 1993). Bargh, Chen and Burrows (1995) present intriguing evidence that even sophisticated social constructs are imbued with motoric representations, to the degree that activating a stereotype will automatically cause motor behaviour typical of that social group.

In some cases, brain imaging techniques have shown the direct involvement of motor areas in 'motor perception' (Stevens, Fonlupt, Shiffrar & Decety, 2000). The existence of 'mirror neurons' also indicates that visual and motor systems share neural circuitry (Gallese, Fadiga, Fogassi & Rizzolatti, 1996). Behavioural experiments by Wohlschläger and Wohlschläger (1998) demonstrated that when subjects mentally rotated a 3D object, performance was slowed if the response used a rotational motor action that was in the opposite direction to the mental rotation. De'Sperati and Stucci (2000) argued that the motor system acts to simulate rotations: in their studies, subjects could more easily judge the rotation of a screwdriver if it was pictured in an orientation easily graspable by their dominant hand. This work is supported by studies that find activation of motor areas during mental rotation tasks (eg. Richter, Somorjai, Summers, Jarmasz, Menon, Gati, Georgopoulos, Tegeler, Ugurbil & Kim, 2000)

Recent research in the stimulus-response compatibility paradigm has found further evidence of the intrusion (or participation) of task-irrelevant motor representations during a perceptual judgement (e.g. Craighero, Fadiga, Rizzolatti & Umilta, 1998). In experiment 1 of Tucker and Ellis (1998) (henceforth T&E) subjects made an orientation judgement (right-side-up/upside-down) about pictures of household objects such as a coffee mug. Each object had an affordance - a handle - on the right or the left side. It was found that subjects were faster when they responded using the hand that was on the same side as the affordance. Later work (Ellis & Tucker, 2000) found a similar compatibility effect when subjects signalled a judgement about an object with a motor action (precision pinch/power grasp) that was appropriate or inappropriate for that object.

This work demonstrates a tight coupling between visual and motor systems: the perception of a graspable object immediately actives a potential motor interaction with that object, even though the affordance is irrelevant to the perceptual judgement. Yet is this just evidence of a rapid link, a transient information hook-up, between visual and motor systems, or is it indicative or a more long term

relationship, whereby the representation of objects is not merely visual, or an amodal list features, but has a motor component that is just as much part of the object bundle.

In Barsalou's (1999a) Perceptual Symbol Systems theory, motor activations such as those revealed in T&E's work should become part of the long term 'simulator' or concept of an object. It has been shown that there are strong associations between object-action pairs (Klatzky, Pelligrino, McClosky & Lederman, 1993) and that object recognition and object grasping utilise overlapping neural networks (Faillenot, Toni, Decety, Gregoire & Jeannerod, 1997). It would be of further interest to show that simply activating a representation of an object causes activation of potential motor interactions. In the first experiment presented here, we extend the results of T&E to show an effect of response compatibility when subjects are recalling a visual image from short term memory.

A considerable body of research demonstrates that conceptual processing has some of the hallmarks of perceptual processing (cf. Barsalou, 1999a). For example, the 'scanning' of mental images mimics the time course of scanning a real image; if a certain location in a mental model is activated, the nearby locations are primed (Bower & Morrow, 1990). If, as the work above suggests, motor activation is often part of perceptual processing, then, we reasoned, perhaps it will be part of conceptual processing as well. The second experiment investigates the role of motor systems in a typical conceptual task – listening to a story.

Experiment 1

We have seen that the presentation of an object's image can activate a potential motor interaction that causally effects a motor response to that stimulus. In this experiment, we would like to see if affordances can be stored in, or reactivated by, short term visual memory.

Method

Subjects 40 undergraduate students of Cornell University participated in exchange for extra credit. All subjects were right handed

Stimuli We compiled 198 images of household objects by using a digital camera and searching public domain image databases. Each picture was in full colour, with a resolution of 500 x 400 pixels. 154 of these were filler images of objects with no obvious affordance, or with an affordance that could be accessed by both the left and right hand equally. We obtained 22 images of objects that had a clear affordance on one side; each of these was mirror reversed. Some stimuli were taken, with kind permission, from Carlson-Radvansky, Covey and Lattanzi (1999). 40 sound files were recorded by an experimenter: the names of the 22 afforded objects, and the names of 18 objects that did not appear in the stimuli set.

Design Each subject was randomly assigned to a response mapping condition. In Left condition, subjects responded 'yes' by pressing the 'S' key and 'no' by the 'K' key. In the Right condition this mapping was reversed.

A schematic of each trial is given in Figure 1. Each trial began with the subjects being reminded of their response mapping. They then pressed the space bar to initiate the trial. Subjects were given a countdown from 3, and then saw 8 images in rapid succession. Each presentation lasted for 200ms. After a 1000ms pause, subjects heard the name of an object. They responded as quickly as possible whether the object was present in the set of 8. No feedback was given.

There were 40 trials. In 18 filler trials, the named object was not present. In the remaining 22 trials, an afforded

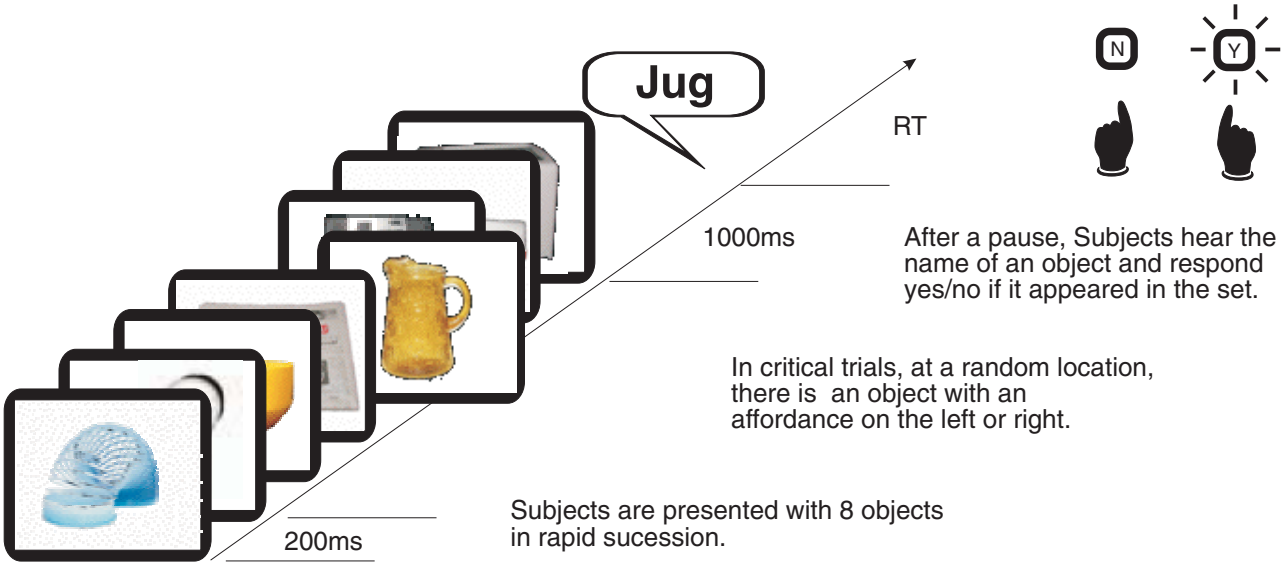


Figure 1: Schematic of Experiment 1.

object was present at a random location, and subsequently named. Half of these objects had an affordance on the left and half on the right, counterbalanced across subjects. The order of the trials and was fully randomised.

Subjects were asked beforehand not to try and name the objects as they appeared. The use of very short presentation times was designed to discourage this verbal labelling strategy. After the experiment., subjects were debriefed and asked to what degree they were able to comply with these instructions.

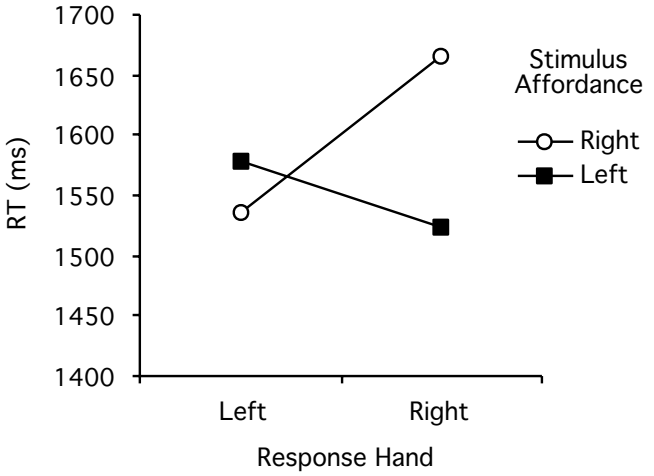


Figure 2 Mean RTs for correctly answered, critical trials, Exp.1

Results In the critical trials, the correct answer was always 'yes'. Accuracy on these trials was 74%. For the remaining analyses, we discarded trials with incorrect responses, and trials with an RT longer than 2.5 standard deviations from the mean (0.006% of the data). Figure 2 shows the trimmed data set.

Subjects making their 'yes' response with the left hand (1551ms) were marginally faster than the those using the right (1601ms), although this effect of response mapping did not approach significance ($F(1,38)=0.20$, $p>.6$). Responses were made slightly quicker when the named object had an affordance on the left (1556ms) versus the right (1598ms), but again this effect did not approach significance ($F(1,38)=1.72$, $p>.19$).

Yet as Figure 2 shows, there was a robust interaction between the stimulus and response conditions that was significant ($F(1,38)=8.22$, $p<.01$). When the hand making the response was on the same side as the affordance of the named object (compatible trials), the mean RT was 1611ms; when the response and object affordance were opposite sides (incompatible trials) the mean RT was almost 90ms faster at 1524ms.

Discussion

Like T&E, we have found an interaction between stimulus affordance orientation and response hand. Yet our results have some interesting differences. First, we have found an incompatibility effect, such that subjects are facilitated when making a judgement using the hand on the *opposite* side to the stimulus affordance. Second, our subjects have response times of about 1500ms, whereas T&E's subjects responded in roughly 700ms. This can be explained by the relative difficulty of the two tasks: our subjects also made about five times as many errors as T&E's.

We suggest that it is primarily this second factor that helps explain the difference in the direction of compatibility. Work by Stoet and Hommel (1999; 2000) shows an interesting time course to the activation and binding of stimulus and response features. These authors use the framework of the Theory of Event Coding, which holds that perception and actions are coded in a common medium. They showed that up until a certain time, stimulus features can be activated such that they facilitate compatible or overlapping responses. However, when the features are activated for a longer time period, they can become bound into an "event file". Once bound, those features are less available for coding compatible responses, and hence an incompatibility effect is found.

This explanation would suggest that if we simply made our visual memory task easier, and hence shortened RTs to the level of T&E's, then we would find a compatibility effect. Our preliminary experiments in this direction are encouraging.

Experiment 2

We have found evidence for an effect of compatibility between a motor response and the affordance of a visual stimulus under conditions where subjects are recalling the relevant object from memory over the course of several seconds. We have seen that there are interesting suggestions form other work that, (a) the direction of the effect hinges upon the time course of feature activation, and (b) these visuo-motor feature codes may be inherent to the representation of functional objects.

If it is the case the motor activation can occur with - perhaps be part of - the activation of object representations, then it should be possible to generate stimulus response compatibility effects from non-visual, verbal descriptions. The difficulty is how to make subjects imagine an afforded object in a particular orientation. Of course, if subjects heard, 'A jug with a handle on the left', then it could be argued that any spatial compatibility effects would be generated by the word 'left' rather than anything related to a motor component of the representation of 'jug.

We attempted to solve this problem by constructing rich scene descriptions. Subjects listened to these stories and then made a yes/no key press in response to a question. The critical trials contained sentences in which the location and orientation of an afforded object was implied by reference to other objects. As in experiment 1, for critical trials this question pertained to the afforded object, and the correct answer was 'yes'. This design allowed us to investigate whether the imaginary orientation of object - indirectly and

verbally described - could interact with the hand used to make a key press.

Pilot work with these stories suggested that subjects made their responses between about 400 and 1800ms – roughly the spanning the RTs of T&E's subjects and our own in Experiment 1. Therefore, we decided to probe the time course of any feature activation by splitting the data at the median RT. We hypothesised that responses below this time would follow a S-R compatibility pattern similar to T&E's, whereas response over the median would have an incompatibility effect resembling Experiment 1.

Method

Subjects 110 right handed Cornell undergraduates participated in this experiment in exchange for course credit. None of the subjects had previously run in Experiment 1.

Stimuli 24 short scene descriptions and questions were written and recorded by the experimenter. The present tense was used throughout. Half of these stories were used as filler items, and the question related to the property of some object or person in the story. The other half of the stories were used as critical trials. Each of these included a description an object with an affordance, and specified the orientation of that object by reference to surrounding items.

First, there was a sentence or two conveying the background scene; in this case, a breakfast table. Two items, one on either side of the scene, were then described. We termed these the 'anchor' objects. In Figure 3 these are a bowl of cornflakes and an egg cup. Then, a third object was mentioned. This was the critical item, an object with an affordance, and was located between the two anchors. Then two phrases specified the orientation of the critical item. They linked a feature or affordance of the critical object with each of the anchors. In the Figure 3 example, the milk jug handle is next to the egg cup, and the spout is pointing towards the bowl. A sentence or two ended the scene description with some further background information. The question was some form of, 'In the center of the [scene] was there a [afforded object]?' The answer to this question on critical trials was, of course, 'yes'.

Figure 3 shows the structure of an example critical trial. This structure allowed us to counterbalance two factors between subjects. Firstly, we varied the left/right positions of the two anchor objects, and hence the afforded object's orientation. Secondly, we switched the order of the two phrases linking afforded object features to anchor objects. This meant that between subjects, we could counterbalance whether a right sided or left sided object was referred to last. Hence any response biases could not be accounted for in terms of simple recency effects.

Design Each subject was randomly assigned to a response mapping condition, as in Experiment 1. Before each trial, subjects were reminded of the response keys to use. There were 24 trials. Each began with a short scene description, about 30 seconds in duration, played over a set

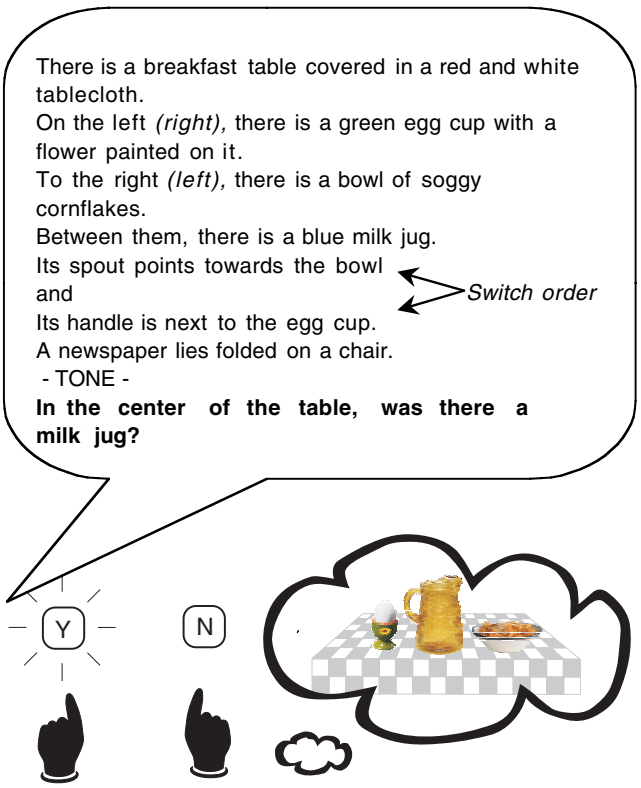


Figure 3: Schematic of a critical item, Exp. 2.

of headphones. At the end of each description, subjects heard a one second tone and then a question concerning the previous information. They were instructed to give their response as quickly and as accurately as possible. Although subjects could take as long as they wanted, if their response time exceeded 5 seconds, that trial was not used in further analysis. We were primarily interested in how subjects represented and manipulated information in order to answer the question, not whether or not they could remember the description verbatim. The cut off point was set on the basis of pilot data to exclude trials in which subjects were struggling to remember details.

Results

The accuracy rate on critical trials was 84.5%. As in Experiment 1, only correct answers to critical trials were analysed. RTs more than 2.5 standard deviations from the mean were excluded from the analysis (4.8% of the data).

The mean RT was 1180ms. The subjects making right handed responses (1158ms) were slightly faster than those using the left hand (1201ms), but this effect did not approach significance ($F(1,108)=0.32$, $p>.5$). In addition, neither the effect of stimulus affordance ($F(1,108)=0.48$, $p>.4$), nor the hand x stimulus interaction ($F(1,108)=0.06$, $p>.8$) approached significance.

To test our hypothesis concerning the time course of stimulus response compatibility effects, we split the trials at the median RT (1020ms) into late and early groups. In order to carry out ANOVAs, we had to remove subjects who did not contribute to both cells. There remained 82 subjects in

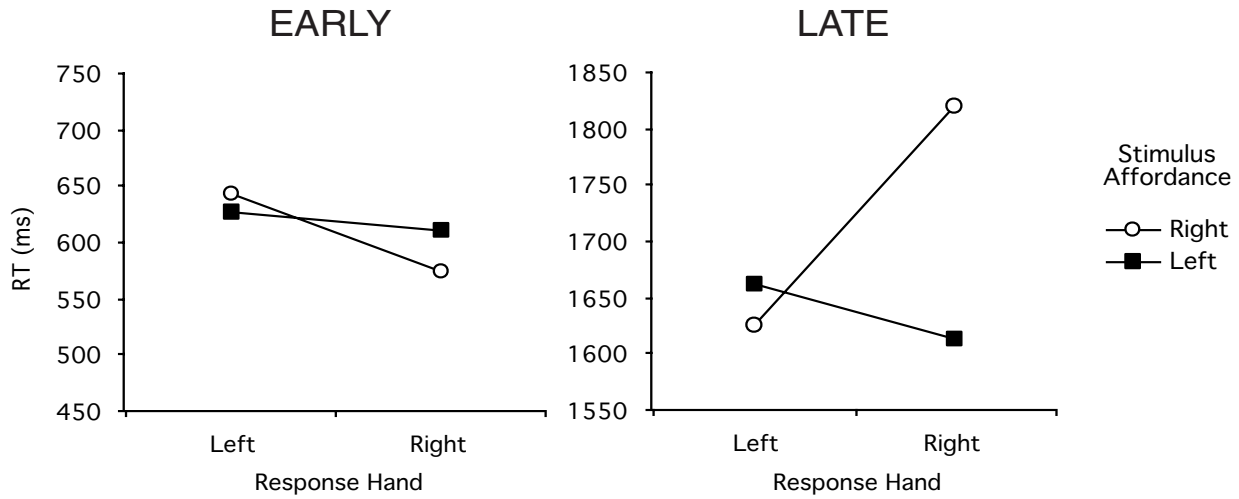


Figure 4: RTs of Experiment 2, split at the median of 1020ms.

the early response condition, and 90 in the late. Their results are shown in Figure 4.

In both groups the main effects of response hand (early, $F(1,80)=2.33$, $p>.1$; late, $F(1,88)=1.12$, $p>.29$). and stimulus affordance (early, $F(1,80)=0.30$, $p>.5$; late, $F(1,88)=0.32$, $p>0.18$) did not approach significance. There was a significant interaction between hand and stimulus in the late group; $F(1,88)=4.38$, $p<.04$. This interaction was not significant in the early responses; $F(1,80)=0.30$, $p>.5$.

Thus, the significant hand X stimulus interaction in the late responses clearly shows an S-R *incompatibility* effect for responses that take place after 1020 ms. In contrast, the early responses show a numerically inverted, albeit nonsignficant, interaction - suggesting an S-R *compatibility* effect for responses that take place before 1021 ms. It is noteworthy that the magnitude of the S-R compatibility advantage for these early responses (20-30 ms) is numerically comparable to that found by T&E.

Finally, in order to get some quantitative indication of whether the interactions taking place in the two time periods are indeed different from one another, we conducted a three-way ANOVA that included early vs. late group as a factor. (Strictly speaking, it is improper treat this factor, which is derived from a dependent variable, as an independent variable. However, the test of a three-way interaction should be sensitive to the relative reaction times across conditions rather than the raw reaction times, and therefore should not be unfairly affected by this procedural irregularity.) As it was necessary to exclude participants who did not contribute to all cells of the design -- many participants were always fast or always slow -- the three-way ANOVA was conducted on the remaining 61 subjects. As predicted, a reliable three-way interaction was obtained where the early reaction times showed a pattern consistent with an S-R compatibility effect and the late reaction times showed a pattern consistent with an S-R incompatibility effect; $F1(1,59)=4.995$, $p<.05$.

Discussion

As hypothesized, the early and late responses show opposite stimulus-response compatibility effects, offering support for the feature activation integration model of Stoet and Hommel (1999). Moreover, we have shown that even in the prime 'disembodied' activity of language comprehension, subjects employ motor representations to construct a mental model, much as Stein's (1994) METATOTO robot created internal maps out of its motor interactions. Thus we have found empirical evidence in support of Bryant's (1998) claim - 'the internal worlds we create do not form maps of external space per se, but of perceptual and behavioral affordances within space.'

General Discussion

Research within the stimulus-response compatibility paradigm has shown that there is a tight coupling between perception and action; indeed, their function is so intimate that it suggests a 'common coding' of perceptual and motor features (Hommel, Müsseler, Aschersleben & Prinz, 2001).

The current experiments reveal one way in which conceptual processes intersect with this tight perception-action arc. Object representations, whether they are memories of a visual stimulus, or part of a mental model generated from a linguistic description, contain motor representations. These results show how motor systems take can take part in 'higher' cognitive functions.

Previous work in our laboratory has shown how *oculo*motor systems participate in the comprehension of spatially extended narratives (Spivey, Tyler, Richardson & Young, 2000), and the spatial indexing of linguistic information (Richardson & Spivey, 2000). Gold and Shadlen (2000) found that in the monkey cortex, competing patterns of activation in populations of the *motor control* neurons will themselves instantiate a 'decision' to saccade. These cases of motor activation occurring as part of a cognitive process are complimented by examples of

linguistic representations intruding upon motor processes. Gentilucci, Benuzzi, Bertolani, Daprati, and Gangitano (2000) showed how automatically activated linguistic representations can intrude upon motor processes. They found that words such as 'near' and 'far' taped on a small wooden bar systematically modulated the kinematics of subjects' reaching behaviour. Moreover, the grammatical class of words, adjectives or adverbs, differentially affected motor control. The results presented in this paper show that the motor system can be modulated not just by spatially extended words, but also by descriptions of afforded objects.

Barsalou (1999b) observed that language is often framed as a means for archiving knowledge. In contrast, he argued for a conception of language comprehension as a 'preparation for situated action'. The involvement of motor representations in memory and mental models we have shown here suggests that language is aptly embodied for this function.

Acknowledgements

The authors wish to thank Bernhard Hommel and Mike Tucker for encouraging and insightful correspondence. Supported by a Sloan Foundation Fellowship in Neuroscience to MJS, and a Sage Fellowship to DCR

References

Bargh, J. A., Chen, M., Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71(2),* 230-244

Barsalou, L.W., (1999a). Perceptual symbol systems. *Behavioral and Brain Sciences, 22(4),* 577-660.

Barsalou, L.W., (1999b). Language comprehension: archival memory or preparation for situated action? *Discourse Processes, 28(1).* 61-80.

Bower, G.H., & Morrow, D.G. (1990). Mental models in narrative comprehension. *Science, 247,* 44-48.

Bryant, D.J., (1998). Human spatial concepts reflect regularities of the physical world and human body. In Oliver & Gapp (Eds.) *Representation and processing of Spatial Expressions*. LEA: London.

Carlson-Radvansky, L.A., Covey, E.S., & Lattanzi, K.M., (1999). What effects on "where": Functional influences on spatial relations. *Psychological Science,10(6),*516-521.

Clark, A. (1997). *Being there: Putting brain, body, and the world together again.* MIT press: Cambridge, Mass.

Craighero, Fadiga, Rizzolatti, Umita. (1998) Visuomotor priming. *Visual Cognition, 5(1/1)*, 109-125.

de'Sperati, C., & Stucchi, N., (2000). Motor imagery and visual event recognition. *Experimental Brain Research, 133,* 273–278.

Ellis, R. & Tucker, M., (2000). Micro-affordance: The potentiation of components of action by seen objects. *British Journal of Psychology 91(4)*, 451-471.

Faillenot, I., Toni, I., Decety, J., Grégoire M., & Jeannerod, M., (1997). Visual pathways for object-orientated action and object recognition: functional anatomy with PET. *Cerebral Cortex*, *7(1),* 77-85.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G., (1996). Action recognition in the premotor cortex. Brain, Volume 119(2), 593-609.

Gentilucci, M., Benuzzi, F., Bertolani, L., Daprati, E., & Gangitano, M. (2000). Language and motor control. *Experimental Brain Research, 133(4),* 468-490.

Gold, J.I., & Shadlen, M.N., (2000). Representation of a perceptual decision in developing oculomotor commands. *Nature, 404(6776*), 390-394

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding: a framework for perception and action planning. *Behavioral and Brain Sciences.* In press.

Iverson, J.M., & Goldin-Meadow, S., (1998). Why people gesture when they speak. *Nature, 396(6708*), 228.

Klatzky, R.L., Pellegrino, J.W., McClosky, B.P., & Lederman, S.J. (1993). Cognitive representations of functional interactions with objects. *Memory and Cognition, 21(3),* 294-303.

Lakoff, G. (1987). *Women, Fire and Dangerous Things.* Chicago and London: University of Chicago Press.

Newton, N. (1998). Review. Being there: putting brain, body and the world together again. *American Journal of Psychology,* 111(1).

Pylyshyn, Z., (2000). Situating vision in the world. *Trends in Cognitive Science, 4(5).* 197-207

Richardson, D.C, & Spivey, M.J., (2000) Representation, space and Hollywood Squares: Looking at things that aren't there anymore. *Cognition, 76(3)* 269-295.

Richter, W., Somorjai, R., Summers, R., Jarmasz, M., Menon, R, Gati, J.S., Georgopoulos, A.P., Tegeler, C., Ugurbil, K., & Kim, S. (2000) Motor area activity during mental rotation studied by time-resolved single-trial fMRI. *Journal of Cognitive Neuroscience,*12(2), 310-320.

Shiffrar, M., & Freyd, J.J., (1993). Timing and apparent motion path choice with human body photographs.. *Psychological Science, 4(6),* 379-384.

Spivey, M.J, Tyler, M., Richardson, D.C. & Young, E., (2000). Eye Movements During Comprehension of Spoken Scene Descriptions. *Proceedings of the Twenty-second Annual Meeting of the Cognitive Science Society*, Erlbaum: Mawhah, NJ.

Stein, L.A., (1994). Imagination and situated cognition. Journal of Experimental and Theoretical Artificial Intelligence, 6.393-407.

Stevens, J.A., Fonlupt, P., Shiffrar, M., & Decety, J., (2000). New aspects of motion perception: Selective neural encoding of apparent human movements.. *Neuroreport: For Rapid Communication of Neuroscience Research, 11(1)*, 109-115.

Stoet & Hommel. (1999) Action planning and the temporal binding of response codes.. *Journal of Experimental Psychology: Human Perception and Performance*, 25(6), 1625-1640.

Tucker, M., & Ellis, R. (1998) On the relations between seen objects and components of potential actions. *Journal of Experimental Psychology: Human Perception and Performance, 24(3),* 830-846

Wohlschläger & Wohlschläger. (1998) Mental and manual rotation. *Journal of Experimental Psychology: Human Perception and Performance, 2(2),* 397-412.

"Language is Spatial": Experimental Evidence for Image Schemas of Concrete and Abstract Verbs

Daniel C. Richardson (dcr18@cornell.edu),
Michael J. Spivey (spivey@cornell.edu)
Shimon Edelman (se37@cornell.edu)
Adam J. Naples (ajn23@cornell.edu)
Department of Psychology, Cornell University
Ithaca, NY 14853 USA

Abstract

Cognitive linguistics and experimental psychology have produced tantalizing hints that a substantial portion of language is encoded in the mind in the form of spatial representations that are grounded in perception and action. Researchers represent these spatial aspects using "image schemas" that depict verbs of motion or spatial prepositions via a 2-D layout of generic icons. In two experiments, we tested naïve subjects' intuitions about such image schemas for concrete action verbs as well as abstract action verbs and psychological predicates. A substantial agreement across subjects was observed in both a forced choice task and a free form computer-based drawing task, for both concrete verbs and abstract verbs. In addition to providing support for the generality of image schemas, the data provide a set of norms for future online studies of spatial representations underlying real-time language processing.

Introduction

Many theorists have argued for a spatial component to language. The arguments are commonly set against an amodal view of representation which defines items in some formal symbolic system. The motivations for proposing an alternative to the symbolic approach range from difficulties in implementing a symbolic system (Barsalou, 1999), commonalties between 'parsing' in the visual system and in language (Landau & Jackendoff, 1993), capturing subtle asymmetries and nuances of linguistic representation in a spatial, schematic way (Langacker, 1987; Talmy, 1983), and a more general account of the mind as an embodied, experiential system (Lakoff, 1987).

If we accept the idea that there is a spatial or perceptual basis to the representation of linguistic items, it would be reasonable to assume that there is some commonality between these representations across different speakers, since by and large we communicate successfully. Therefore, we might expect that there would be a consensus among subjects when we ask them to draw simple diagrams representing words. Theorists such as Langacker (1987) have produced large bodies of diagrammatic linguistic representations, arguing that they are constrained by linguistic observations and intuitions in the same way that 'well formedness' judgements inform more traditional linguistic theories. However, it remains to be seen whether naïve subjects share these intuitions and forms of representation. Therefore, in the same way that psycholinguists use norming studies to support claims of preference for certain grammatical structures, we propose to survey a large number of subjects and see if there is a consensus amongst their spatial representations of words.

Recent work has also documented the mapping between spatial linguistic terms and the mental representation of space (eg Hayward & Tarr, 1995; Carlson-Radvansky, Covey & Lattanzi, 1999; Schober, 1995). Although there are consistencies in the ways in which spatial language is produced and comprehended (eg Hayward & Tarr, 1995), the exact mapping appears to be modulated by such factors as visual context (Spivey-Knowlton, Tanenhaus, Eberhard & Sedivy, 1998), the common ground between conversants (Schober, 1995) and the functional attributes of the objects being described (Carlson-Radvansky et al., 1999).

When language refers directly to explicit spatial properties, locations, and relationships in the world, it is quite natural to expect those linguistic representations to have at least some degree of overlap in their format. Spatial language terms appear to be grounded, at least somewhat, in perceptual (rather than amodal) formats of representation. However, an important component of the work presented herein involves testing for this representational format in an arena of language that does *not* exhibit any literal spatial properties: abstract verbs (such as 'respect' and 'succeed'). Much work in cognitive linguistics has in fact argued that many linguistic and conceptual representations (even abstract ones) are based on metaphorical connections to spatially laid out "image schemas" (Gibbs, 1996; Lakoff, 1987; Langacker, 1987; Talmy, 1983). This work suggests that if consistency across subjects is observed for spatial depictions of *concrete* verbs, then one should also expect such consistency for *abstract* verbs. Experimental evidence for this kind of broad consensus among speakers would extend the "language is spatial" hypothesis beyond spatial terms, and make some experimentally supported, albeit preliminary, claims about abstract language as well.

There are various old and new results suggesting that there is some consistency among speakers in the *visual imagery* associated with certain ideas and concepts. For example, Scheerer and Lyons (1957) asked subjects to

match the referents ‘gold’, ‘silver’, and ‘iron’ with three drawings which had previously been produced by other naive subjects. At least one set of these drawings (which resembled sine, saw tooth, and square waves, respectively), were correctly matched by 85% of the subjects. Lakoff (1987) offers anecdotal evidence that when asked to describe their image of an idiom such as 'keeping at arms length' people have a considerable degree of commonality in their responses, including details such as the angle of the protagonist's hand. Similarly, Gibbs, Strom and Spivey-Knowlton (1997) carried out empirical work querying subjects about their mental images of proverbs such as 'a rolling stone gathers no moss' and found a surprising degree of agreement – even about fine details such as the stone bouncing slightly as it rolled. Experimental work has shown that the listing the features of a concept involves something akin to visually inspecting its properties (cf Barsalou, 1999).

This approach extends beyond the simple visual properties of a concept, towards more schematic or spatial representations. Barsalou’s (1999) perceptual symbol system theory endorses the view held by several theorists (e.g. Lakoff, 1987; Gibbs, 1996) that to some degree abstract concepts are represented by a metaphoric relation to more concrete domains. For example, it is argued that the concept of ‘anger’ draws on a concrete representation of ‘liquid in a container under pressure’. There is some debate over how central these metaphorical aspects are to the representation of abstract concepts. But for the representation of time, at least, there is strong experimental evidence that subjects’ reasoning is structured by a metaphorical relation to space.

Boroditsky (1999) observed that English speakers tend to use horizontal spatial metaphors when talking about time, whereas Mandarin speakers use both horizontal and vertical. In a reaction time study, speakers from both languages were asked true/false questions about time (e.g. ‘March comes earlier than April’) It was found that Mandarin speakers responded faster when they had been presented with vertical rather than horizontal spatial primes, and the reverse was true for English speakers. The result is particularly impressive since both groups carried out the experiment in English. Boroditsky (2000) identified two schemas that are used in both spatial and temporal domains: ego moving and time/object moving. She found that the use of one type of schema in the spatial domain, would prime a judgement of the same schema in a temporal domain.

In this paper, we empirically test the claim that between subjects there is a coherence to the imagistic aspects of their linguistic representations. To this end we will address two questions – *Do subjects agree with each other about the spatial component of different verbs?* And, *Across a forced choice and an open ended response task, are the same spatial representations being accessed?* It would be of further interest if the subjects’ diagrams bore resemblance to those proposed by theorists such as Langacker (1987). However, as with more standard norming studies, the real value of the data will be in allowing us to generate prototypical representations that could be used as stimuli for studies of online language comprehension.

Experiment 1

Methods

Subjects 173 Cornell undergraduates participated in exchange for course credit.

Design We selected 30 verbs to fill out a concreteness by spatial layout, 2x3 factor design. Using the MRC psycholinguistic database, we divided the words into high and low concreteness. These two concreteness groups were each divided into 3 groups based on the expected primary axes of their image schemas (vertical, horizontal, and neutral), based on our survey of the cognitive grammar literature. This 2x3 factor design was filled with a list of 30 verbs. Each was placed in the past tense in the form of a simple rebus sentence, with circle and square symbols representing agents and patients.

The subjects were presented with a single page, containing a list of the verbs and four pictures, labelled A to D. Each one contained a circle and a square aligned along a vertical or horizontal axis, connected by an arrow pointing up, down, left or right. Since we didn't expect any interesting item variation between left or right placement of the circle or square, the horizontal schemas differed only in the direction of the arrow.

For each sentence, subjects were asked to select one of the four sparse images that best depicted the event described by the sentence (Figure 1)

The items were randomised in three different orders, and crossed with two different orderings of the images. The six lists were then distributed randomly to subjects.

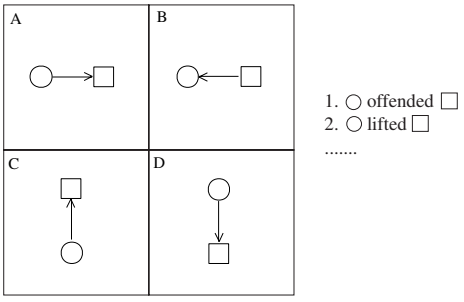


Figure 1: Example of the questionnaire in Experiment 1.

Results

Subjects’ responses are summarised in Table 2. The most frequently chosen image column is in bold for each verb. On average, for any given verb, the particular image orientation that was most popular was chosen by 63% of the subjects. The second most popular was chosen by 21%, the third by

10% and the fourth by 5%. This suggests a substantial degree of agreement between subjects.

To test our predictions concerning the primary axes of the verbs' image schemas, we converted the forced choice data into axis angles. The left and right image schemas were assigned an angle of 0, and the up and down image schemas a value of 90. See Table 2.

A two-way ANOVA by-items analysis revealed a significant main effect of expected axis ($F(2,24)=30.30$, $p<0.0001$), and the effect of concreteness did not approach significance ($F(1,24)=1.84$, $p>0.18$). There was, however, a significant interaction ($F(2,24)=5.28$, $p<0.02$), indicating that the effect of expected axis was more dramatic for concrete verbs than for abstract verbs. Planned comparisons revealed that, even among the abstract verbs, the mean axis angle of the expected-horizontal verbs was lower than that of the neutral verbs and the mean axis angle of the vertical verbs was greater than that of the neutral verbs (all *p*s<.05).

Table 1: Percentage of subjects choosing each image

Concreteness	Expected Axis	Verb	Up	Down	Left	Right
HIGH	Horizontal	fled	7.2	4.2	**80.8**	7.8
		pointed at	7.2	3.6	0	**89.2**
		pulled	6	5.4	**75.4**	13.2
		pushed	7.2	3.6	1.2	**88**
		walked	9	3.6	24	**62.9**
	Neutral	hunted	9.6	20.4	1.8	**68.3**
		impacted	7.2	37.1	3	**52.7**
		perched	12	**76**	6.6	5.4
		showed	15	9	10.2	**65.9**
		smashed	3.6	**66.5**	1.2	28.7
	Vertical	bombed	4.8	**86.8**	1.8	6.6
		flew	37.7	**44.3**	15	3
		floated	32.9	**56.3**	7.8	3
		lifted	**87.4**	9.6	2.4	0.6
		sank	22.2	**71.9**	4.2	1.8
LOW	Horizontal	argued with	11.4	13.8	12.6	**62.3**
		gave to	8.4	9.6	1.2	**80.8**
		offended	9	31.7	24.6	**34.7**
		rushed	10.2	10.8	23.4	**55.1**
		warned	10.8	22.2	6	**61.1**
	Neutral	owned	5.4	**55.7**	18.6	20.4
		regretted	19.8	24	**41.3**	15
		rested	14.4	36.5	**40.1**	9
		tempted	16.8	11.4	**45.5**	26.3
		wanted	15.6	7.8	15.6	**61.1**
	Vertical	hoped	**45.5**	15.6	7.2	31.7
		increased	**73.7**	7.2	9.6	9
		obeyed	22.8	4.2	**64.7**	8.4
		respected	**53.9**	3	14.4	28.7
		succeeded	**40.1**	35.9	10.8	13.2
		Means	20.9	26.2	19	33.8

Table 2: Mean Axis angle.

Expected Axis / Concreteness	Horizontal	Neutral	Vertical
High	10	46	82
Low	25	37	55

Discussion

It appears that there is a considerable degree of agreement between subjects. This consistency was seen in both concrete verbs of motion, eg 'lifted', and abstract verbs, such as 'respected'. Yet it could be argued that this coherence mainly reflects the artificial and limited nature of the forced choice ask, rather than a commonality of deeper significance between subjects' representations. In our next experiment, we allowed subjects to create their own image schemas in an open response task.

Experiment 2

In this experiment, we asked subjects to create their own representation of the sentences using a simple computer based drawing environment.

Method

Subjects Twenty-four Cornell University undergraduates participated in exchange for course credit. None of these subjects had participated in Experiment 1.

Design Subjects were presented at random with a sentence from Experiment 1. They were given as much time as they required to draw a schematic representation of the sentence. When they had finished, they clicked a done button and were given the next sentence.

The drawing environment is shown in Figure 2. Subjects could drag the shapes on to central canvas. Any number of shapes could be used, and they could be repositioned. Subjects could also use up to 3 arrows. By holding down modifier keys, the arrows could be re-sized and rotated.

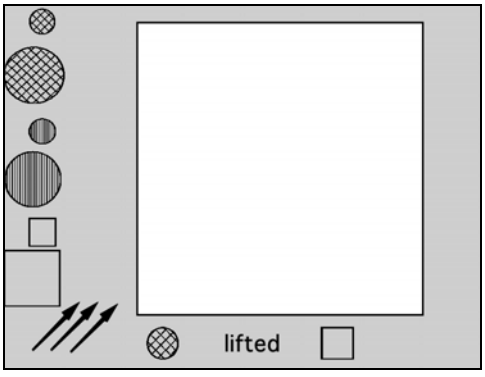


Figure 2: Screen shot from Experiment 2.

Results

Subjects spent approximately a minute completing each drawing. Some subjects produced quite sparse, schematic representations; others attempted more complex depictions. Figures 3 and 4 show a random selection of drawings of the concrete verb 'argued with' and the more abstract verb 'respected'.

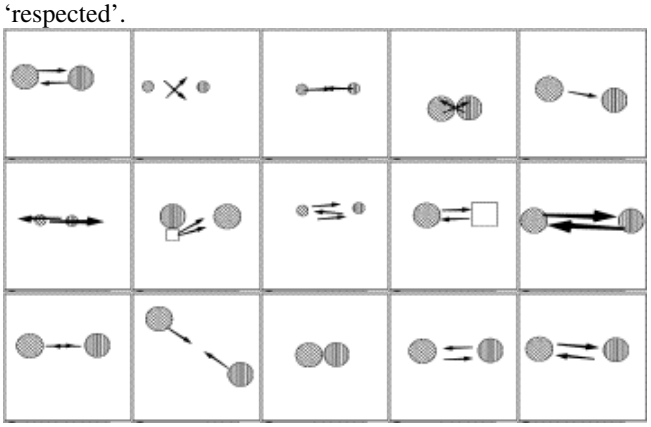


Figure 3: Example depictions of "ARGUED WITH".

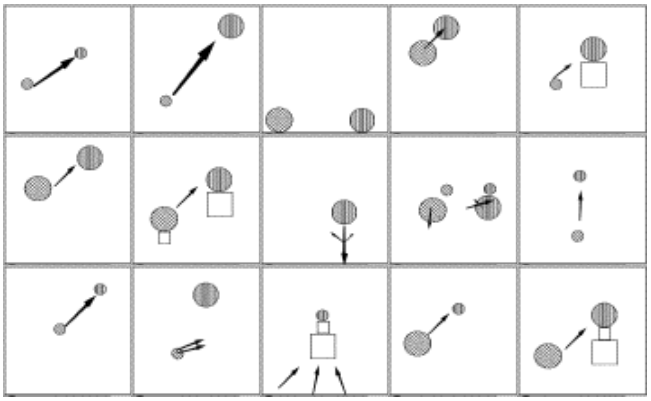


Figure 4: Example depictions of "RESPECTED".

The majority of subjects appeared to represent the verbs schematically using quite sparse images. However, there were a few subjects who, despite the limitations of the drawing toolbox, attempted to *pictorially* represent the verbs. For example, in the third and fourth figures in the second row of Figure 4, we can see that the subjects have drawn humanoid figures, using the arrows as arms. Indeed, since they were the only items that could be rotated and resized, the arrows were often used as generic lines forming a pictorial drawing. For this reason, we decided to ignore the arrows in our analysis, and focus on the relative positions of objects.

Using the coordinates of objects in the drawing, we defined the 'aspect angle', as a value between 0 and 90 which reflects the horizontal versus vertical extent of each drawing. If one imagines a box drawn around the center points of all objects in a picture, the aspect angle is the angle of a diagonal line connecting the lower-left and upper-right corners of the box. If the objects are aligned on a horizontal axis, the aspect angle would be 0; on a vertical axis, 90.

Note that the aspect angle collapses left-right and top-bottom mirror reflections of a drawing. We decided to use this measure since we were primarily interested in the horizontal versus vertical aspect of each drawing. In addition, the initial starting orientation of the arrows (Figure 2) might bias subject towards a right rather than left, and an upwards rather than downwards layout in their drawings: this bias would be avoided in calculating the aspect angle. Figure 5 graphically represents the aspect angle data in what we have termed a 'radar' plot. Each verb's mean aspect angle (solid line) is shown together with its standard error (shaded fan area), and included is the mean axis angle of that verb in the forced choice task of Experiment 1 (dashed line). The means for each condition are shown in the final column of Figure 5.

These results were subjected to a ANOVA items analysis. The only significant effect was of expected axis ($F(2,24)=6.69, p<0.005$). The mean aspect angle for the horizontal group was 21˚, neutral 36˚ and vertical 45˚.

Discussion

Despite the free form nature of the task, it seems that there was a reasonably high degree of agreement between subjects. The mean standard error for all verbs was 6.5 degrees. Previous work has found that subjects consistently place the flow of action from left to right when depicting events (Chatterjee, 2001), but our subjects employed contrasting horizontal and vertical image schemas as well. Moreover, there is considerable consistency between the drawings and the results of the forced choice task. We observed a significant correlation between the mean aspect angles for the verbs in the two tasks (R= 0.71).

General Discussion

We have presented data that suggest there is an impressive degree of coherence in the spatial, schematic components of some linguistic representations. Two different tasks attempted to tap these representations. A forced choice task with very sparse images appeared to produce comparable results to a creative, open response task. In this sense, the data suggest a positive answer to two of our opening questions - Do subjects agree with each other? and Do the two tasks assess the same underlying representations?

We would also argue that our third question - *Do naive subjects agree with trained linguistic intuitions?* – can be given a qualified 'yes'. In both experiments, the expected axis had a significant effect on the orientation of subjects' responses. Figure 5 reveals some informative cases where our expectations were defeated by subjects. For example, in our neutral condition, both 'perched' and 'rested' were consistently given a vertical aspect angle by subjects in both tasks. This observation highlights the importance of using normative methodologies to accompany traditional linguistic methodologies.

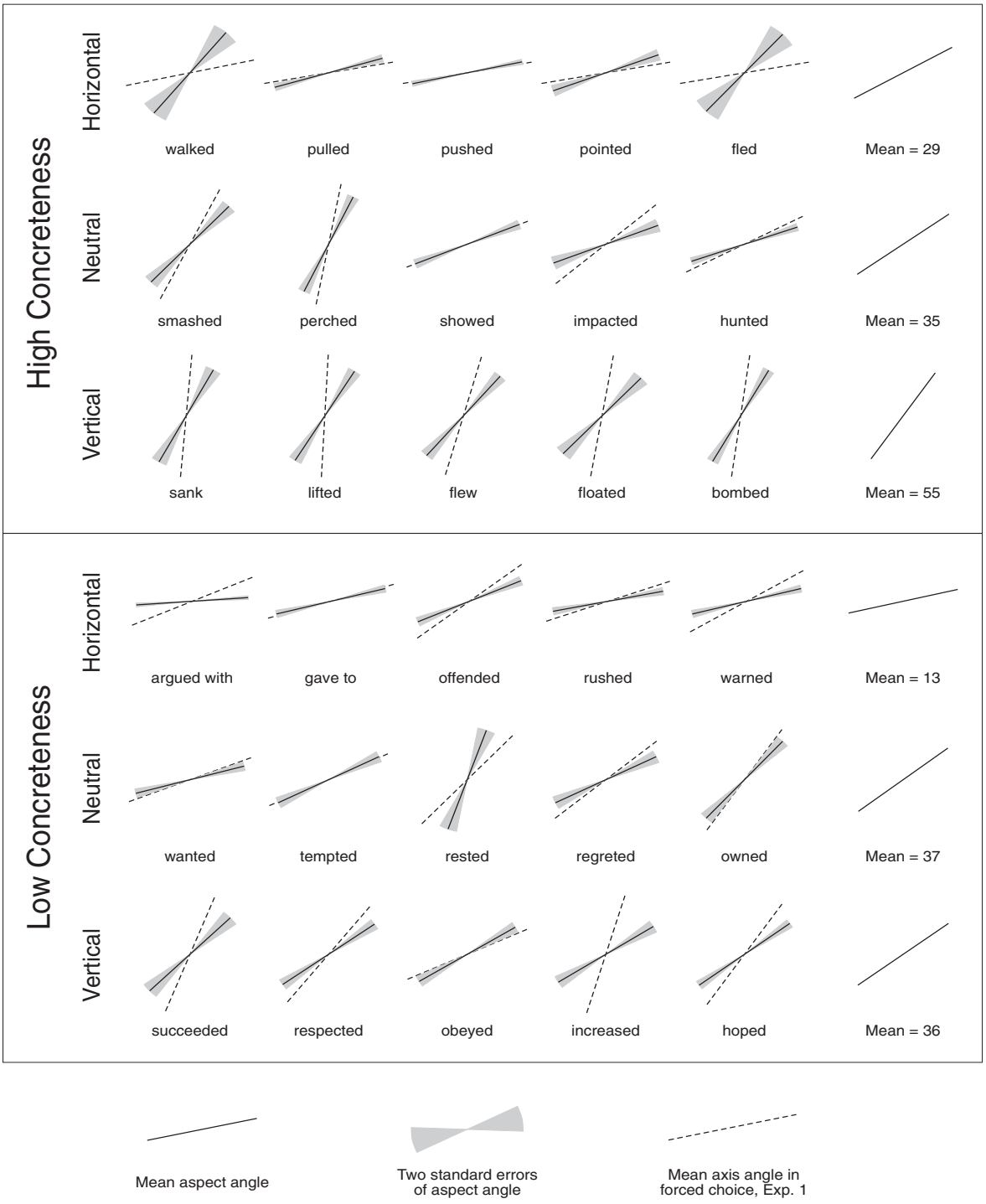


Figure 5: 'Radar' plots of mean aspect angles in subjects' drawings, Exp.2

Although these norming studies demonstrate considerable (and perhaps surprising) agreement between the intuitions of cognitive linguists and naïve subjects, we would argue that the true value of these results is in the predictions they generate for real-time language processing. Just as offline word similarity ratings predict online performance in word priming tasks, we hope that our offline data will predict effects of spatial priming for online language comprehension. In this way, we can further test whether the spatial formats of linguistic representation suggested by these results are indeed fundamental components of language processing in natural situations, and not just artefacts of contemplative metalinguistic intuitions induced only by unusual offline tasks.

There are theoretical grounds for proposing such experiments. For example, in Barsalou's (1999) perceptual symbols systems, a simulator "controls attention across the simulation" (p.604). If our experiments have successfully tapped the spatial element of such simulators, concepts, or image schemas, then we would expect linguistic processing to modulate spatial attention in some manner. For example, if it is the case that the representation of certain words have a spatial element with some degree of verticality, then perhaps priming subjects with a vertical image schema such as those used in Experiment 1 would facilitate a lexical decision task for words such as 'respect' and 'succeed'.

In addition to more standard psycholinguistic paradigms, we hope to use eye movement data to investigate the spatial component of linguistic processing. It has been well demonstrated that mental imagery and mental models exhibit properties of an analog spatial layout (eg Denis & Cocude, 1992; Bower & Morrow, 1990). Work in our laboratory has demonstrated that this spatial component is evidenced in subjects' eye movements. When passively listening to a scene description and staring at a blank wall, subjects tend to make eye movements that correspond to the direction of the events described (Spivey, Tyler, Richardson & Young, 2000).

Similarly, Kaden, Wapner and Werner (1955) showed that visually perceived eye level is influenced by the spatial components of words. Subjects sat in a dark room and saw luminescent words at their objective eye level. Subjects then had the words moved up or down, until they were at the subjective eye level. Words with an upward connotation ('climbing', 'raising') had to be placed lower to be perceived as being at eye level, whereas words with a downward component ('falling', 'plunging') had to be placed above the objective eye level.

We hope that these image schema norms will allow us to measure spatial effects at a finer grain of representation than previous eye movement studies. Given that corresponding eye movements are made during an explicitly spatial description, perhaps they will also be made during an implicitly spatial description (due to metaphorical connections to image schemas). For example, we might find a bias towards vertical eye movements when listening to a description of John's respect for Mary.

Many of these findings are, or will be, surprising. If the purpose of communication is to *guide attention* and *coordinate action*, one should expect visuo-spatial information to play a causal role in linguistic processing. Future work will determine how strong a role.

Acknowledgements

We would like to thank Jamie Cheung for her assistance with data collection, and Dick Neisser for valuable discussions. Supported by a Sloan Foundation Fellowship in Neuroscience to MJS, and a Sage Fellowship to DCR.

References

Barsalou, L. W., (1999). Perceptual symbol systems. Behavioral and Brain Sciences,, 22(4) 577-660.

Boroditsky, L., (1999).First-language thinking for second language understanding: Mandarin and English speakers' conception of time. *Proceedings of theTwenty-first Annual Meeting of the Cognitive Science Society*, Erlbaum: Mawhah, NJ

Boroditsky, L., (2000) Metaphoric structuring: understanding time through spatial metaphors *Cognition*, *7,* 1-28.

Bower, G.H., & Morrow, D.G. (1990). Mental models in narrative comprehension. *Science, 247,* 44-48.

Carlson-Radvansky, L. A., Covey, E S., Lattanzi, K. M., (1999). What effects on "where": Functional influences on spatial relations. *Psychological Science, 10(6),* 516-521.

Chatterjee, A., (2001). Language and space: some interactions. *Trends in Cognitive Sciences, 5(2),* 55-61.

Denis, M., Cocude, M., (1992). Structural properties of visual images constructed from poorly or well-structured verbal descriptions *Memory and Cognition 20(5).*497-506

Gibbs, R. W., (1996). Why many concepts are metaphorical *Cognition, 61*, 309-319.

Gibbs, R. W., Strom, L. K., Spivey-Knowlton, M. J., (1997). Conceptual metaphors in mental imagery for proverbs. *Journal of Mental Imagery*, *21(3-4),* 83-109.

Hayward, W.G., Tarr, M.J., (1995). Spatial language and spatial representation. *Cognition*, *55(1),* 39-84.

Kaden, S.E., Wapner, S., & Werner, H., (1955) Studies in physiongomic perception: II. Effect of directional dynamics of pictured objects and of words on the position of the apparent horizon. *Journal of Psychology, 39,* 61-70.

Lakoff. G., (1987). *Women, Fire and dangerous things*. The University of Chicago Press: Chicago.

Landau, B., Jackendoff, R., (1993). "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, *16,* 217-265.

Langacker. R.W., (1987). An introduction to cognitive grammar. *Cognitive Science, 10(1),* 1-40.

Scheerer, M., & Lyons, J., (1957) Line drawings and matching responses to words. *Journal of Personality, 25*, 251-273.

Schober, M.F., (1995). Speakers, addressees, and frames of reference: whose effort is minimized in conversations about locations? *Discourse Processes*, *20,*219-247.

Spivey, M.J, Tyler, M., Richardson, D.C. & Young, E., (2000). Eye movements during comprehension of spoken scene descriptions. *Proceedings of the Twenty-second Annual Meeting of the Cognitive Science Society*, Erlbaum: Mawhah, NJ

Spivey-Knowlton, M.J., Tanenhaus, M., Eberhard, K., & Sedivy, J., (1998). Integration of visuospatial and linguistic information: language comprehension in real time and real space. In Oliver & Gapp (eds) *Representation and processing of Spatial Expressions*. LEA: London.

Talmy. L., (1983). How language structures space. In Pick and Acredolo (eds), *Spatial orintation: theory, resarch and application.* Plenum Press: New York.

Efficacious Logic Instruction: People Are Not Irremediably Poor Deductive Reasoners

Kelsey J. Rinella (rinelk@rpi.edu)
Department of Philosophy, Psychology & Cognitive Science
The Minds & Machines Laboratory
Rensselaer Polytechnic Institute (RPI)
Troy, NY 12180 USA

Selmer Bringsjord (selmer@rpi.edu)
Department of Philosophy, Psychology & Cognitive Science
Department of Computer Science
The Minds & Machines Laboratory
Rensselaer Polytechnic Institute (RPI)
Troy, NY 12180 USA

Yingrui Yang (yangyri@rpi.edu)
Department of Philosophy, Psychology & Cognitive Science
The Minds & Machines Laboratory
Rensselaer Polytechnic Institute (RPI)
Troy, NY 12180 USA

Abstract

Cheng and Holyoak, and many others in psychology and cognitive science, subscribe to the view that humans have little context-independent deductive capacity, and that they can't acquire such a capacity through instruction in formal logic. This position is based, in no small part, upon C&H's well-known investigation of the efficacy of an undergraduate course in logic in improving performance on problems related to Wason's Selection Task, in which they found the benefit of such training to be minimal (Cheng, Holyoak, Nisbett, & Oliver, 1986). We believe, based on the encouraging results of a new study involving a similar pre-test/post-test design on a logic class at RPI, that the results obtained in the Cheng & Holyoak study serve to highlight problems with the way logic has historically been taught (related to techniques unavailable or impractical before the advent of heavy computer saturation in higher education), rather than to suggest that humans are unable to learn to reason. This prompted the reevaluation of conclusions based on C&H's research, requiring a new theory of meta-reasoning, Mental MetaLogic.

Introduction

The backlash against Piaget's claims (e.g., see the claims in Inhelder and Piaget, 1958) that humans naturally acquire competence in (elementary extensional) logic has "ruled the roost" in the psychology of reasoning for some time. Recently there has been some thought that perhaps the inherent irrationality of the species has been exaggerated (see Bringsjord, Noel, & Bringsjord, 1998; Evans & Over, 1996; Rips, 1994). This article is targeted specifically at the claims made by Cheng et. al. (1986) that not only are humans inherently bad at logic, but we are unable through training in formal logic to learn how to reason in abstract, context-independent fashion. One of the experiments they report, experiment 2, involves a pre-test/post-test design in which students in a logic class are tested on their understanding of how the conditional works in examples—the improvement they report is minimal. Using the same design, but a different instructional method, our results indicated a significantly greater improvement.

The three major reasons put forth in this presentation that the logic class at RPI differed from those in previous studies is that it taught disproofs, diagrammatic techniques, and, "Rigorous and general-purpose procedures for formalizing natural language logic problems in first-order logic so that they can then be solved by automated theorem provers". (For more on this last technique, see Bringsjord & Ferrucci, 2000.) Briefly, disproofs are proofs that one sentence does not follow from a set (possibly empty) of givens. Put another way, they are proofs that, given whatever premises one has, it is not possible to prove the goal, nor is it possible to prove the negation of the goal. The software used in the course, HYPERPROOF (Barwise

and Etchemendy, 1994), allows students to see how sentential information in first-order logic interacts with a toy world which acts as the domain of discourse (were it not for the existence of this world, it would not be possible to perform disproofs in a way remotely similar to that used in the RPI course) through experimentation and practice problems. Because of this, our students learn the meanings of the sentences in a much more understandable fashion while retaining the abstractness and universality of formal logic—these are the diagrammatic techniques. Finally, one of the most challenging tasks involved in solving many of the problems presented by psychologists of reasoning is finding the intended content in the words presented. The translation procedures mentioned above allow students to make fewer errors on these sorts of tasks. The virtues of these advances are discussed in some detail in Bringsjord, Noel, & Bringsjord (1998), with additional data presented in Bringsjord & Rinella (1999).

To demonstrate the abovementioned diagrammatic techniques from HYPERPROOF, which may be unfamiliar to many readers, consider figure 1:

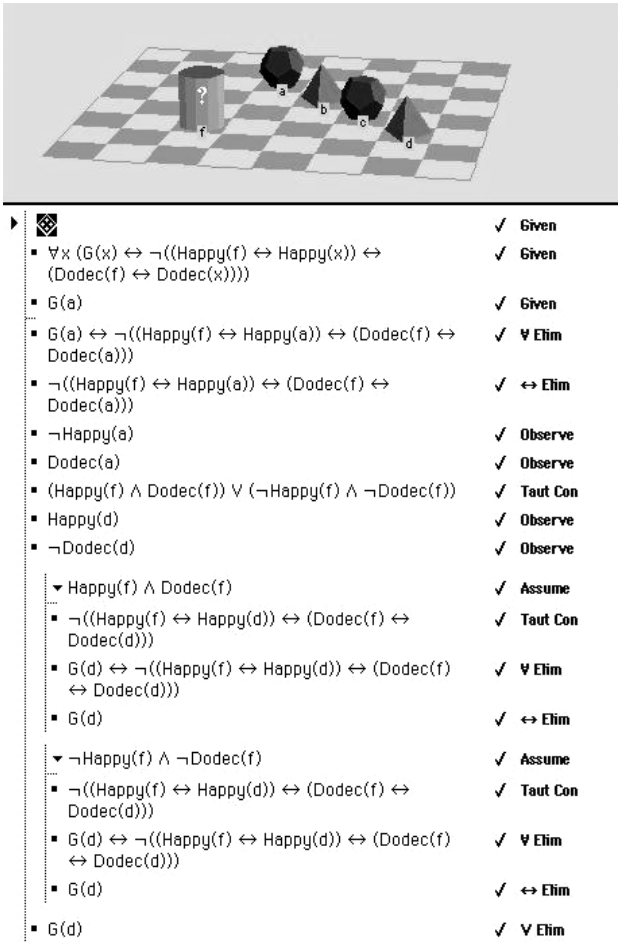


Figure 1:The THOG Problem

The HYPERPROOF window consists of two major areas: on top, there is the toy world, which shows the locations and properties of a small number of objects (in this case, five); on the bottom, sentential logic inferences proceed toward the goal. This particular example is a formalization of the THOG problem, a common problem used in the psychology of reasoning. The first given sentence claims that an object has the property G if and only if it has either the emotional state (happy or unhappy) or the shape (dodecahedron or tetrahedron) of the object f, but not both. Since we don't know f's shape (HYPERPROOF hides shape when it is not know by placing a cylindrical box over the object, which is why f appears to be a cylinder) or emotional state, we need the information in the second given, that object a has property G, to determine which of the other objects has property G. The proof proceeds by first manipulating the givens to extract the information that object a has either the emotional state or shape of f, but not both. Unlike regular sentential logic, we are then able to observe from the world the status of a, noting that it is unhappy and a dodecahedron (in some problems, it is actually necessary to use information from the sentential section to add information to the world, often allowing the user to detect an object's location, shape, or size, so this information moves up to the world as well as down to the sentences). From this, we infer that f must be either a happy dodecahedron or an unhappy object of a different shape. Finally, we show that, in either of these instances, object d also has property G, and conclude by stating that d must have G. Students using this system have the advantage that they are able to see what the sentences mean—rather than proceeding merely by manipulating the symbols of the sentences according to rules they have learned by rote, they begin to understand how different configurations of objects alter the effects of different sentences.

Method

We gave students enrolled in Rensselaer Polytechnic Institute's Introduction to Logic class in the Fall term of 1998 a pre-test including Wason's Selection Task as problem one, the THOG problem as problem two, and five other problems from previous work by psychologists of reasoning or from experience with tests of logic encountered by students in other contexts (e.g., two of the problems were straightforward adaptations of problems the Board of Regents of New York State say every New York high school students should be able to solve; Verzoni & Swan, 1995). A similar test, mathematically matched for problem type and difficulty, was given as a post-test appended to the final exam. Though there is insufficient space here to present them, the complete pre-test and post-test are available online:

http://www.rpi.edu/~faheyj2/SB/INTLOG/pre-test.f98.pdf, http://www.rpi.edu/~faheyj2/SB/INTLOG/post-test.f98.pdf. An example pair of THOG-like problems follows. In both cases, of course, students had to provide correct justifications.

> **2**
> Suppose that there are four possible kinds of objects:
>
> - an unhappy dodecahedron
> - a happy dodecahedron
> - an unhappy cube
> - a happy cube
>
> Suppose as well that I have written down on a hidden piece of paper one of the attitudes (unhappy or happy) and one of the shapes (dodecahedron or cube). Now read the following rule carefully:
>
> - An object is a GOKE if and only if it has either the attitude I have written down, or the shape I have written down, but not both.
>
> I will tell you that the unhappy dodecahedron is a GOKE. Which of the other objects, if any, is a GOKE?

The analogous problem on the post-test was the following:

> **2**
> Suppose that there are four possible kinds of objects:
>
> - an smart tetrahedron
> - a stupid tetrahedron
> - a smart cube
> - a stupid cube
>
> Suppose as well that I have written down on a hidden piece of paper one of the mental attributes (smart/stupid) and one of the shapes (tetrahedron/cube). Now read the following rule carefully:
>
> - An object is a LOKE if and only if it has neither the mental attribute I have written down, nor the shape I have written down.
>
> I will tell you that the stupid tetrahedron is a LOKE. Which of the other objects, if any, is a LOKE?

Note that a direct, unreflective transfer of reasoning brought to bear on the first of these problems to the second won't yield a solution to the second. This pair of problems (and this holds true for each pair on our pre-test/post-test combination) will not "match" at the surface level in English, nor at such a level in the propositional calculus. However, we needed pairs of problems that, at the level of proof or disproof, could be said to be very similar, formally speaking. Without such mathematical similarity we wouldn't be able to justifiably say that the problems, from the standpoint of formal deduction, are essentially the same. Figure 1 above presents a proof-theoretic solution to the first of the THOG-like problems—it is at this level of detail that difficulty must be matched without allowing the same argument form to work a second time.

Subjects who took only one of the two tests were discarded, to ensure that every participant had exposure to the entire course, leaving exactly 100 participants. After the first test, we abandoned asking the subjects whether they had seen the question before. There were two reasons for this: prior experience did not correlate with success on the questions, and the problems on the post-test were so similar in theme and difficulty that it was very likely that their experience with the pre-test would generate false positive responses. We also of course asked for justifications for their answers, hoping that out of the data we would be able to divine an appropriate scheme for categorizing the unstructured and heterogeneous responses we were likely to get.

As a preface to the first test, we gathered some biographical information, including the sex of the participants, the location of their high school, and previous logic experience. Since New York's Board of Regents has decreed that students must learn logic in their math courses, we hypothesized that attending high school in New York state would increase performance on tests of reasoning. We also hypothesized that previous experience in logic would increase scores on the pre-test, but that this effect would be reduced or eliminated by the post-test.

Results & Discussion

As expected, the averages on the pre-test were significantly lower than on the post-test, 3.89 correct compared to 5.11 correct. A paired-samples t-test reported an extremely low (t = -8.393) probability of no effect, suggesting that taking the logic class did improve students' ability to reason logically. Full results for each of the questions appear in table 1, below:

Table 1: Individual Question Results

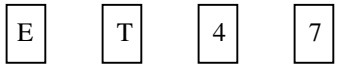
	Test 1	Test 2	t	Significance
Question 1	29	84	-9.563	0.000
Question 2	72	83	-2.076	0.040

Question 3	77	94	-3.597	0.001
Question 4	55	80	-4.639	0.000
Question 5	90	98	-2.934	0.004
Question 6	7	58	-9.768	0.000
Question 7	59	14	7.595	0.000

Though the improvement was significant at the .01 level for each of the first six questions except question two (which had a problem with a ceiling effect, but was still significant at the .05 level), there were three questions that particularly attracted our attention. Questions one and six showed extremely low initial rates of success, but great improvement—this suggests that these question types may be particularly amenable to improvement by instruction in formal logic. Question seven totally reversed our expectations—students did markedly worse on the post-test.

Individual Question Findings

The first result of some import is the comparison of Wason's Selection Task and its analogue (in each case, question one) on the post-test. These problems were chosen to test the ability of students to comprehend the use of the conditional in a context-free setting. The difficulty our subjects had on the pre-test with this problem very much agrees with the performance of Cheng & Holyoak's participants on their pre-test (1985). From this poor performance, and the lack of improvement, Cheng & Holyoak concluded that people are not good at using the conditional in a context-independent manner. On the pre-test, the problem looked like this:

> **1**
> Suppose that I have a pack of cards each of which has a letter written on one side and a number written on the other side. Suppose in addition that I claim the following rule is true:
>
> - If a card has a vowel on one side, then it has an even number on the other side.
>
> Imagine that I now show you four cards from the pack:
>
> 
>
>
> Which card or cards should you turn over in order to decide whether the rule is true or false?

The analogous problem (from Verzoni & Swan, 1995) on the post-test follows:

> **1**
> Suppose that you are doing an experiment for a biology expedition. You learn before starting on this expedition that insects can be one of two kinds, a spade fly or a bevel wasp, and that insect color is either black or green. Your task is to study insects in order to find out if a certain rule is false. The rule is:
>
> - If an insect is a spade fly, then it is black.
>
> You see an insect that is green. Which of the following would be true about the insect if it violates the rule?
>
> **a** The insect is a spade fly.
> **b** The insect is a bevel wasp
> **c** The type of insect does not matter.

Because these are the problems which are identical in underlying form to those used by Cheng et. al. in the aforementioned 1986 study, we were quite pleased to discover that our methods had induced an improvement from 29 correct responses to 84 correct responses, an extremely impressive improvement. This confirms our initial hypothesis, and allows our results to very directly be compared with previous work.

The second question which drew our attention because of the extremely poor (well below chance) performance on the pre-test. Since this was the question relating to *reductio ad absurdum,* or proof by contradiction, which is an integral part of the work our students do with HYPERPROOF during the semester. Such a proof, from the standpoint of the psychology of reasoning (which focuses on untrained reasoning), is exotic, but from the standpoint of mathematics and mathematical logic, it's thoroughly routine, and is therefore part and parcel of an introductory course of the type we offered. The full text of this question from the pre-test follows:

> **6**
> We will use lower-case Roman letters *a, b, c, …* to represent propositions. Let the symbol '¬' stand for 'it is not the case that.' Let the symbol '∨' stand for 'or.' Let the symbol '→' stand for 'if-then', so that $p \rightarrow q$ means 'if p then q.'
> Given the statements
> $\neg\neg c$
> $c \rightarrow a$
> $\neg a \vee b$
> $b \rightarrow d$
> $\neg(d \vee e)$
> which one of the following statements must also be true? (Check the correct answer.)
>
> $\neg c$

e
h
¬*a*
all of the above

Once again, of course, we gave a corresponding problem on the post-test. Alert readers will have realized that the answer to **6** is "all of the above," which of course means that *h* must be true given the quintet. The reason for this, of course, is that the quintet is inconsistent, and therefore a straightforward proof for *h* (or any other propositional variable) can be easily given.

The final question of particular interest was question seven, the results from which seemed to suggest that our course had made students worse at reasoning of this type. It involved fairly complex reasoning on statements which were presented in English, thus requiring more effort to extract meaning. Looking for an explanation, we noticed that the following sentence appeared in the pre-test version of this question (from Smullyan, 1982), "'At least one of them did [tell the truth],' replied the Dormouse, who then fell asleep for the rest of the trial." The question from the post-test, which was intended to be analogous, included the following sentence, which was supposed to play the same role, "'Well, one of them did [tell the truth],' replied Devin, who then fainted and remained unconscious for the remainder of the investigation." This difference seemed potentially problematic.

On further investigation, we noticed that many of the justifications on the second problem suggested that subjects were having problems interpreting this statement by Dr. Devin, "Well, one of them did." This can be (and was) interpreted in two common ways, as, "One and only one of them did," or as, "At least one of them did." If these are both appealing interpretations, as they seemed to be for many of the participants, there is no entirely logical way to figure out the answer. A very small number of particularly clever subjects assumed that there would be enough information given in the question to figure out the right answer, and realized that one of the interpretations, that only one of them told the truth, did not fulfill this requirement. These students then rejected this option and solved the problem. However, doing all of that, which seems to be the only way other than guessing that subjects were able to correctly answer the problem, is far more difficult than interpreting the analogous statement in the missing jam problem, which was made by the Dormouse, in response to questioning about whether the Mad Hatter and March Hare had spoken the truth: "At least one of them did." Since this is clearly much more explicit, we have considered the seventh questions on the two tests to be sufficiently different that they are no longer appropriate for comparison. Unfortunately, it was not possible to counter-balance the pre-test and post-test, because of the high degree of availability of students to each other; to make both sets of questions available in this way would have introduced an unacceptably strong confound.

Demographic Data

Without the last question on each test, averages dropped to 3.30 correct on the pre-test and 4.97 on the post-test. This improved the value of the t-statistic to t= -13.653. This indicates even more clearly that subjects did in fact improve their ability to succeed on tests of reasoning due to the instruction in logic, and that the improvement was of a fairly substantial magnitude.

Interpreting the justifications turned out to be fairly problematic. Our initial attempt was a fairly subjective rating system based on the opinions of a competent logician, but there is a potentially very important confound in this method, which is that a correct answer is much more likely to suggest to a reader that the subject knew what s/he was doing, even if this is somewhat underdetermined by the written justification. Since we are most interested in the correlation between justification quality and success rate, this rating system was unacceptable. However, the information from the justifications did turn out to be useful in checking to make sure that the questions were interpreted as we intended, and further exploration may reveal a more objective way to code this data such that it may be made more useful.

Point-biserial correlations (appropriate for categorical data of this type, rather than more common values) were calculated between sex, high school attendance in New York state/elsewhere, and previous logic experience and the two test averages, and similar Pearson correlations calculated within those two groups of factors. Nothing significant came out of sex. Surprisingly, we did not observe a significant correlation between high school state and performance on either test. Previous logic experience did correlate positively with performance on the pre-test, but not on the post-test, as expected, with Pearson correlations of .246 (significant at the 0.05 level) and -.021, respectively. This indicates that the course did make up for any disadvantage less experienced students may have had coming in, and also suggests that performance on the pre-test was actually higher than it ought to have been, because we assumed that incoming students would not have been formally trained in logic. Since only one-tenth of the subjects were so trained, and sometimes in courses that dealt only tangentially with logic, we suspect this effect was negligible.

Unsurprisingly, none of sex, high school state, and previous logic experience correlated with each other. Also unremarkable was the highly significant correlation of .465 found between the pre-test and post-test scores—subjects with higher initial ability are likely to have higher ability after the end of the course.

Conclusions

The proposition that humans are unable to learn to reason better through instruction in formal logic seems to be disconfirmed by these data. This naturally does not mean that pragmatic effects hold no power over our attempts to use what deductive competence we have developed, nor does it suggest that all tests of reasoning will show improvement following an arbitrarily-selected course in logic. However, Cheng and Holyoak's proposed pragmatic reasoning schema theory (see Cheng et. al., 1986; Cheng and Holyoak, 1985; and Holyoak and Cheng, 1995) needs revision to remain a plausible candidate explanation of human reasoning. Yang and Bringsjord (2001, 2001) have suggested an alternative theory of human and, by extension, machine reasoning, viz., Mental Metalogic (MML), which allows pragmatic reasoning schemas to continue to play a role in human cognition, but not alone. In MML, mental models and mental logic exist side-by-side with such schemas, and a higher-level choice mechanism selects the most appropriate form of reasoning for the task at hand. In this regard, it's important to note that MML draws from a part of mathematical logic hitherto untapped in cognitive science: metatheory.

Recent advances in the teaching of logic (particularly HYPERPROOF) were utilized in the course used in the study, and this may help explain the differences in the results seen by Cheng and company, and those found in our study (for a mathematical analysis of HYPERPROOF in the context of "heterogeneous" reasoning consistent with Mental MetaLogic, see Barwise & Etchemendy 1995). In addition to the technological sophistication and concomitant improvement in available techniques, our interest in matters related to the psychology of reasoning may help to explain these results. In the class at RPI, students were encouraged to think about problems from the standpoint of metatheory: to ponder the way that they might approach logic problems (e.g., from the standpoint of searching for proofs *a la* mental logic, or from the standpoint of disproofs and mental models.) We routinely presented several options and contrasted their power, and also studied the reasoning process itself. The increased introspection about the reasoning process that this may have produced in our students is another factor which distinguishes the RPI logic class from previous subjects of similar experiments.

We believe that the reason standard logic instruction has not improved performance on tests of the sort given by proponents of the pragmatic reasoning schema theory may be related to the importance of one or more of the factors we have mentioned, which are historically missing in most classes. If this is correct, contra Cheng and Holyoak, it is not the level of abstraction that keeps logic instruction from being efficacious in improving reasoning. With the right theoretical perspective (MML), and pedagogical techniques which recognize the efficacy of non-pragmatic reasoning associated with that perspective, students can easily carry out difficult context-independent deduction suggestive of that of a professional logician or mathematician.

References

Barwise, J., & Etchemendy, J. (1994). *HYPERPROOF*. Stanford, CA: CLSI.

Barwise, J. & Etchemendy, J. (1995) Heterogeneous Logic. In *Diagrammatic Reasoning*, Glasgow, J., Narayanan, N.H., and Chandrasekaran, B., eds. Cambridge, MA: MIT Press.

Bringsjord, S. & Ferrucci, D. (2000) *Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus, A Storytelling Machine*. Mahwah, NJ: Lawrence Erlbaum.

Bringsjord, S., & Rinella, K. Hard Data in Defense of Logical Minds. *Annual International Conference on Computing and Philosophy*. Carnegie-Mellon University, August 6, 1999.

Bringsjord, S., Noel, R., & Bringsjord, E. (1998). In Defense of Logical Minds. *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 173-178). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cheng, P. W., Holyoak, K. J., Nisbett, R. E., & Oliver, L. M. (1986). Pragmatic versus Syntactic Approaches to Training Deductive Reasoning. *Cognitive Psychology*. 18, 293-328.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic versus Syntactic Approaches to Training Deductive Reasoning. *Cognitive Psychology*. 17, 391-416.

Evans, J., & Over, D. E. (1996). *Rationality and Reasoning*. Hove, East Sussex, UK: Psychology Press.

Holyoak, K. J., & Cheng, P. W. (1995). Pragmatic Reasoning About Human Voluntary Action: Evidence from Wason's Selection Task. In S. E. Newstead & J. Evans (Eds.), *Perspectives on Thinking and Reasoning*. Englewood Cliffs, NJ: Lawrence Erlbaum Associates.

Inhelder, B., & Piaget, J. (1958). *The Growth of Logical Thinking from Childhood to Adolescence*. New York, NY: Basic Books.

Smuyllan, R. (1982). *Alice in Puzzleland*. New York, NY: Morrow.

Rips, Lance. (1994). *The Psychology of Proof*. Cambridge, MA: MIT Press.

Verzoni, K. & Swan, K. (1995) On the Nature and Development of Conditional Reasoning in Early Adolescence. *Applied Cognitive Psychology*. 9, 213-234.

Yang, Y., & Bringsjord, S. (2001). Mental Possible World Mechanism and Logical Reasoning in GRE. (under submission).

Yang, Y., & Bringsjord, S. (2001). Mental MetaLogic: A New Paradigm in Psychology of Reasoning. (under submission).

Using cognitive models to guide instructional design: The case of fraction division

Bethany Rittle-Johnson (br2e@andrew.cmu.edu)
Kenneth R. Koedinger (koedinger@cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University
Pittsburgh, Pa 15213 USA

Abstract

Cognitive modeling can be used to compare alternative instructional strategies and to guide the design of curriculum materials. We modeled two alternative strategies for fraction division, and the models led to specific empirical predictions of the benefits and drawbacks of each strategy. These insights provided concrete suggestions for developing lessons on fraction division, including a new potential strategy that combines the benefits of the two strategies. This on-going work illustrates the potential of cognitive modeling for informing the design of better mathematics curricula.

Background

Although U.S. students are fairly proficient at performing routine calculations, they lack a conceptual understanding of mathematics and have difficulty solving non-routine problems (Lindquist, 1989; Jakwerth, 1999). These findings have spurred many educators to call for an increasing focus on building understanding and problem solving skill in mathematics instruction. The National Council for Teachers of Mathematics (NCTM) standards state the overarching learning goal as: "Students must learn mathematics with understanding, actively building new knowledge from experience and prior knowledge," (p. 16, NCTM, 2000). The standards proposed by NCTM, and curriculum and evaluations based on them, have met with opposition from advocates of "back-to-basics" approach (Mathematically Correct, 2000). Mathematicians, politicians, teachers and parents have raised concerns that students are not learning their arithmetic facts and basic computational skills and are lobbying to abandon these reform efforts. Further, many teachers have been resistant to changing their teaching practices and doubt the benefits of reform-based curricula.

Although there is agreement on the need to improve the mathematics curriculum in the U.S., there is considerable disagreement on how the curriculum should be changed. Controversy over how to teach fraction division helps to illustrate this fundamental conflict over whether the curriculum should focus on gaining conceptual understanding or on proficiency in retrieving facts and executing computational procedures. To solve fraction division problems, students are traditionally taught the computational procedure of inverting the divisor and changing the operation to multiplication (invert-and-multiply strategy). As an alternative, the NCTM 2000 standards proposed a picture division strategy where students draw a picture of the starting amount, repeatedly "cut off" groups of the size specified by the divisor and count the resulting number of groups. For example, to solve six divided by 3/4, students could draw a line six units long, divide each unit into fourths, and then start at six and mark off groups of three fourths to find how many 3/4 are in six. According to the standards: "Lacking an understanding of the underlying rationale [for invert and multiply], many students are therefore unable to repair their errors and clear up their confusions about division of fractions.. Carefully sequenced experiences with [picture division] problems such as these can help students build an understanding of division of fractions." (p. 218 NCTM, 2000).

In principle, an ideal approach to informing the debate on how best to teach a particular mathematical topic is to conduct a multi-year, multi-site experimental study comparing a reform-based approach with a back-to-basics approach. In addition to the practical limitations of this approach, an empirical evaluation does not explain the reasons for the results or offer insights into how to apply these results to other topics.

We have begun to explore the role of cognitive models in helping to inform this debate. Developing cognitive models offers four key advantages. First, developing cognitive models requires precise and unambiguous specification of problem representations and action sequences and allows for detailed comparisons of problem solving strategies. Second, the specificity of the models leads to generation of specific hypotheses that can be tested through smaller, focused, empirical studies. Third, cognitive models can be used to understand and explain empirical results, allowing researchers to understand the mechanisms underlying the differences and to extrapolate the findings to other domains. Finally, inspection and evaluation of these models yield concrete suggestions for better content and methods of teaching a particular topic.

To illustrate the potential of cognitive modeling for informing the current debate in mathematics instruction, we describe our use of cognitive modeling to guide the design of lessons on fraction division (as part of a middle-school math curriculum we are developing). Rational number concepts and procedures are a cornerstone of middle-school mathematics, but U.S. students perform poorly on a range of rational number problems, including fraction division problems. (e.g. Lindquist, 1989; Lesh & Landau, 1983). Fraction division is a representative topic in mathematics for which the standards-based and back-to-basic movements have proposed alternative strategies.

In the current paper, we present cognitive models of both fraction division strategies, outline predictions for learning

and transfer that are revealed by the models, and offer preliminary implications of the models for instructional design, including a new strategy for fraction division that was suggested by this work.

Cognitive Models

Our cognitive models are based on ACT-R theory, which breaks knowledge into two main categories – a declarative knowledge base of facts and a procedural knowledge base of production rules (Anderson, 1993). Declarative knowledge includes both prior domain knowledge and representation of the current problem situation. A production rule is a simple IF-THEN statement that manipulates declarative knowledge, and a series of production rules model actions for solving a problem.

Model of Picture Division Strategy

We collected informal verbal protocols from five sixth-grade students while they solved basic fraction division problems such as 15 ÷ 1 1/2. The students had received no formal instruction on fraction division. One student spontaneously used a picture division strategy, and the other students were provided with a picture and encouraged to try using the strategy.

The combination of the task analysis and students' think alouds revealed 4 main sub-goals for implementing this strategy: 1) identify the values in the problems and draw the appropriate picture, 2) mark the picture into the appropriate size groups, 3) count the number of groups, and 4) convert the remainder (if there is one) to a fractional value. These four sub-goals translate into 13 key steps or actions that the problem solver must take (see Table 1; a dotted line designates the beginning of a new sub-goal). These actions were instantiated as productions in an intelligent tutoring system that is based on ACT-R theory (Anderson, Corbett, Koedinger & Pelletier, 1995).

Table 1: Cognitive model of picture division strategy

Productions	Student Example
1. Identify-starting-amount	Here's her 8 foot long board
2. Draw-whole-starting-amount	[draws line with 8 sections]
OR	
3. Draw-mixed-starting-amount	
4. Identify-size-of-groups	And she wants each one [shelf] to be a half,
5. Identify-value-of-divisions	So, I'd split each one in half
6. Draw-divisions	[marks each whole in half]
7. Identify-step-size	[group size = 1; skip to 9]
8. Mark-first-group	
9. Mark-next-group	
10. Finished-marking-groups	
11. Count-whole-groups	And then that's how many shelves. [counts] 16.
12. Identify-remaining-divisions	NA
13. Step-size-as-denominator-of-remainder	NA

Note: Extra productions would be needed to solve problems where the denominators of the dividend and divisor are different and one is not a multiple of the other (e.g. 3/5 ÷ 1/3).

Table 2: Example declarative knowledge chunk: Representation of 2/3

```
QUANTITY 2/3>
  isa number
  whole          0
  top-number     2
  bottom-number 3
  parts-per-whole 3  ;;Picture Division strategy only
  needed-parts    2  ;;Picture Division strategy only
```

The productions in the picture division model rely on meaningful representation of problem information in declarative memory. First, selection of this strategy comes from representing the meaning of division as finding the number of groups of a given size in the starting amount. Second, the productions rely on a quantity-based representation of fractions. Students need to represent fractions as parts of a whole (e.g. 2/3 is two out of three equal size parts) rather than only as a visual arrangement of numbers (e.g. 2 is the top number and 3 is the bottom number). Table 2 provides a sample declarative chunk used in representing the problem.

Model of Invert-and-Multiply-Strategy

The invert-and-multiply strategy can be broken into 4 main sub-goals: 1) identify the dividend and the divisor, 2) if needed, convert whole numbers and mixed numbers to fractions, 3) invert the divisor and multiply the two fractions, and 4) if needed, simplify the answer by

Table 3: Cognitive model for invert-and-multiply strategy

Productions	Student Example
1. Identify-dividend	12
2. Identify-divisor	1 1/2
3. Whole-dividend-to-fraction	12/1
4. Identify-mixed-dividend	NA
5. Identify-mixed-divisor	
6. Mixed-to-fraction	1 1/2 is 3/2
7. Invert-divisor	So, make it 12/1 * 2/3
8. Multiply-top-&-btm-#s	Equals 24/3
9. Improper-to-mixed	That is 8
10. ID-whole-#-answer	[Done]
11. ID-if-quotient-is-reducible	NA
12. Reduce Fraction	NA

converting an improper fraction to a mixed number and/or by putting the fraction in simplest terms. These four sub-goals translate into 11 key steps or actions that the problem solver must take (see Table 3; dotted lines designate the beginning of a new sub-goal).

The declarative knowledge used by the invert-and-multiply strategy differs from that used by the picture division strategy. In the invert and multiply strategy, division is represented as performing actions on numbers. Fractions are represented as a visual arrangement of digits, and quantity-based knowledge is not used (see Table 2).

Empirical Support for Our Models

We designed a brief intervention to validate and refine our cognitive models. The students had already been taught the invert-and-multiply strategy, and we gave them a brief lesson on the picture division strategy. Thirty-two ninth-grade students from two math classes for below-average math students participated in the study.

On the first day of our study, students received a 10-minute lesson on fraction division from their classroom teacher. The teacher discussed how to solve two types of fraction division problems using each strategy. Instruction on the picture division strategy focused on forming a quantity-based representation of the problem without detailed instruction on the actions (productions) for implementing the strategy. The teacher then reviewed the steps for using the invert-and-multiply. After this brief lesson, students were randomly assigned to use one of the two strategies to solve a set of problems. Students solved two problems using the assigned strategy and received feedback and help in finding the correct answer if needed. Students then solved a set of 10 problems without feedback or help – 4 instructed problems (problems with whole numbers and/or unit fractions), 5 transfer problems (problems with non-unit fractions and with mixed numbers), and a fraction multiplication problem. Students had approximately 20 minutes to solve the problems, and their compliance with the strategy instructions was high. In the invert-and-multiply group, there was no trace of students using a pictorial strategy, and in the picture division group, a picture was drawn on 83% of attempted problems. Four days later, students were asked to solve a parallel set of 10 problems using any strategy they wanted.

Students had difficulty learning the picture division strategy from our brief lesson. On Day 1, they solved 60% of the instructed problem types correctly, but only 9% of the transfer problems correctly. Many students got stuck and did not finish the assessment; students only attempted 57% of the problems (compared to students attempting 96% of problems in the invert-and-multiply group).

Not surprisingly, students who were assigned to use the familiar invert-and-multiply strategy solved more problems correctly on Day 1, compared to the Picture Division group (49% vs. 28% correct; $F(1,30) = 16.96$, $p < .01$). They solved 89% of the instructed problem types correctly, but only 31% of the transfer problems correctly (although they had previously received instruction on these problem types as well.) Student had particular difficulty when the problems involved mixed numbers. Only half of the students solved at least one problem of this type correctly.

When students were free to choose any strategy on Day 2, students used the more familiar and well practiced invert-and –multiply strategy on a majority of problems (M = 62% of trials with a mean accuracy of 60%). The picture division strategy was used on 10% of problems, and only by students who were assigned this strategy on Day 1.

To help explain the difficulties of each strategy, students' incorrect solutions were classified using the productions in the relevant cognitive model. We distinguished between failing to initiate a production and implementing a production incorrectly (an error). The ease of coding student solutions is an additional benefit of developing cognitive models. We report solution data from Day 1, but a similar pattern arises on Day 2.

Tables 4 and 5 show the distribution of failures and errors over the productions for each strategy. On the picture division strategy, students often did not know how to start the problem. When students attempted the problem, they often did not succeed on the first sub-goal – identifying values and setting up the picture. There were a surprising number of errors in identifying the dividend and in drawing the divisions correctly (e.g. students added 3 extra divisions per whole for 1/3, thus making fourths). Students' errors on identify-parts-per-whole varied by problem type, suggesting that an additional production was needed in our model. When the dividend was a fraction, students sometimes divided the fractional amount, rather than the whole, into the specified number of parts (e.g. for 1/2 ÷ 1/10, dividing the half into 10 sections). Students need an extra production for mapping the parts-per-whole to the parts-per-fraction (e.g. if 10 division in one whole, half as many (5) in 1/2). We have very little data on the difficulty of productions that occur later in the sequence because students often abandoned this strategy.

Students using the invert-and-multiply strategy were much more likely to attempt to solve a problem, and the majority of mistakes arose from failing to or incorrectly converting mixed numbers to fractions. However, this error did not cause students to abandon the strategy. Rather, students made illegitimate adaptations to the strategy, such as inverting the fractional portion of the divisor and then multiplying the whole number portions and the fraction portions separately (e.g. 8 2/3 ÷ 2 1/3 = 8 2/3 * 2 3/1 = 16 6/3). Further, students' errors on the fraction multiplication problem suggested that the conditions for firing the invert-divisor production were overly general for many students. Half of the students in the invert and multiply group inverted the second fraction before multiplying.

Table 4: Classification of students' incorrect solutions using the picture division strategy on day 1

Action/Production	No. of Errors	No. of Failures
<Start Problem>	NA	64
ID /Draw-whole-starting-amt	6	1
ID /Draw-mixed-starting-amt	7	-
Identify-size-of-groups	-	-
Identify-value-of-divisions	6	11
Draw-divisions	10	-
Identify-step-size	7	1
Mark-first-group	-	-
Mark-next-group	1	-
Finished-marking-groups	-	-
Count-whole-groups	2	1
Identify-remaining-divisions	NA	NA
Step-size-as-denominator-of-remainder	NA	NA

Table 5: Classification of students' incorrect solutions using the invert and multiply strategy on day 1

Action/Production	No. of Errors	No. of Failures
<Start problem>	NA	8
ID Dividend	-	-
ID Divisor	-	-
Whole-to-fraction	-	-
ID-mixed-dividend & Mixed-to-fraction	13	15
ID-mixed-divisor & Mixed-to-fraction	4	9
Invert-divisor	2	9
Multiply-top-&-btm-#s	4	8
Improper-to-mixed	3	9
ID-if-common-factor	NA	NA
Reduce Fraction	NA	NA

Implications of the results for the cognitive models

The empirical results revealed a necessary refinement to the picture division model and validated the other productions in the two models. Students' errors when using the picture division strategy indicated that an additional production was needed when the dividend was a fraction. Otherwise, the models captured students' behaviors quite well.

The empirical results also provide information on common buggy rules and on the frequency of correct productions "failing to fire". Students' buggy rules will be modeled as production rules, allowing us to identify the source of the differences in the correct and incorrect productions. This information can be used to target instruction at addressing or preventing these errors.

These results also highlight the importance of the declarative knowledge structures. The ninth-grade students in this study did not seem to form quantity-based representations of fractions or to represent division as finding the number of groups of a certain size in the starting amount. Without these declarative knowledge structures, students had great difficulty implementing the initial productions for the picture division strategy. In contrast, the invert-and-multiply strategy only relies on a superficial representation of division and of the position of the digits in fractions, although a quantity-based representation of the values could be used to recognize errors in its execution (e.g. that multiplying the whole numbers will lead to too large of an answer). After more than 5 years of instruction on the division operator and on fractions, these students did not seem to be forming meaningful representations of either.

Predictions from the models

Cognitive models of the picture division and invert-and-multiply strategies can lead to comparative predictions for 1) difficulty of learning each strategy, 2) efficiency of using each strategy once learned, 3) generality of each strategy to the range of fraction division problems, 4) retention of the strategies, and 5) transfer.

First, the ease of learning the two strategies depends on students' prior knowledge. In particular, learning difficulty should be predicted by two factors – how students represent fractions and division and how well they know symbol manipulation rules for working with fractions. If students form quantity-based representations of fractions and attach meaning to the division operation, learning the picture division strategy should be relatively straightforward since a majority of the productions are based on familiar and well practiced knowledge (e.g. marking sections and counting). However, if students only represent fractions as visual arrangements of digits and division as manipulating symbols, this representation is not compatible with the strategy, so the strategy will be difficult to learn. The invert-and-multiply is not dependent on a quantity-based representation of fractions. In contrast, the ease of learning this strategy depends on how well students already know productions for converting whole and mixed numbers to fractions and for converting improper fractions to mixed numbers.

Second, our models support the predictions that the two strategies will not be equally efficient once they are mastered. Although the total number of productions to learn is similar in the two strategies (13 vs. 12), the number of production firings is often higher for the picture division strategy because some productions must fire many times. For example, to solve $6 \div 3/4$, the draw-divisions production fires 18 times and the mark-next-group production fires 6 times. Thus, to solve this problem, the picture division strategy has 32 production firings whereas the invert-and-multiply strategy has 6 production firings. On a majority of problems, the invert-and-multiply strategy is more efficient than the picture division strategy once the strategy is mastered.

Third, the ease of applying the two strategies to the full range of fraction division problems is not equivalent. Once the full set of productions is mastered for the invert-and-

multiply strategy, it can be applied to any fraction division problem. In contrast, the picture division strategy becomes very cumbersome if the dividend is large, the denominator of the divisor is large, or if the denominators of the dividend and divisor are not "friendly" (i.e. one denominator is not a factor of the other, such as 3 and 5). The first two constraints require an unmanageable number of firings of the draw-divisions and mark-next-group productions. The third constraint requires a new set of productions for finding equivalent fractions with a common denominator, thus necessitating extra productions that are not well grounded in the situation. Overall, the picture division and invert-and-multiply strategy can both be used to solve a majority of fraction division problems, but the invert-and-multiply strategy has the advantage of more uniform difficulty on all types of problems.

The fourth prediction concerns the retention of the two strategies and confers an advantage to the picture division strategy. In ACT-R, recall is based on spreading activation, so knowledge that is connected to a richer network of knowledge chunks is easier to recall (Anderson, 1993). The picture division strategy utilizes rich knowledge representations of quantities and operations, so this network of relations should facilitate recall. In contrast, the invert-and-multiply strategy utilizes sparse, visual-based representations that are not connected to a rich knowledge base, so this strategy should be harder to recall after a delay. Our results indicate that students have difficulty correctly retrieving all of the relevant productions for invert-and-multiply. In addition, both level-of-processing and dual-code theories of memory (Craik & Lockhart, 1972; Paivio, 1971) suggest that the richer problem representations utilized by the picture division strategy should lead to better recall of this strategy, compared to the invert-and-multiply strategy. Thus, we predict that recall of the picture division strategy will be more robust.

Fifth, the models lead to very different transfer predictions. Inspection of the models indicates no overlap in the productions that are used by each strategy, so learning one strategy will not aid learning of the other. The two strategies also transfer differently to other topics. When knowledge chunks are activated, their memory trace is strengthened (Anderson, 1993), so quantity-based representations of fractions and a meaning-based representation of division are strengthened (and possibly refined) when students use the picture division strategy. Thus, learning the picture division strategy should facilitate performance on tasks utilizing these representations. Representing fractions as part-whole quantities provides a powerful declarative knowledge structure for choosing and implementing a variety of strategies for tasks such as comparing magnitudes, estimating, or adding and subtracting fractions. The picture division strategy should also transfer to decimal division since it strengthens a meaningful representation of division and many of the productions can be used to solve division problems with decimals. In contrast, the invert-and-multiply strategy should facilitate performance on problems involving other fraction operations or algebraic simplification. This strategy strengthens productions that are also used for adding, subtracting and multiplying fractions, such as converting improper fractions to mixed numbers, reducing fractions, and multiplying fractions (although students may over-generalize the strategy and also invert the second fraction when multiplying fractions). Productions from this strategy can also be applied to simplifying algebraic expressions. Overall, the two strategies should aid performance on very different types of transfer problems.

Developing cognitive models of the two strategies leads to precise predictions of the benefits and drawbacks to each strategy. The picture division strategy should be easy to learn if students have quantity-based representations of fractions, should be recalled after a delay, and should transfer to tasks such as comparing fractions and dividing by a decimal. In contrast, the invert-and-multiply strategy should be easy to learn if students already know productions for manipulating fractions, should be efficient and broadly applicable once mastered, and should transfer to other fraction operations and to algebra.

Implications for Instructional Design

Comparing the benefits and drawbacks of each strategy allows for an informed decision on whether and how to teach each strategy. Neither of the strategies was strong along all five dimensions that we considered (difficulty of learning, efficiency, generality, retention and transfer). Instead, there were trade-offs for learning each strategy.

How the fraction division problems are represented in declarative memory helps to explain the benefits and drawbacks to each strategy. The picture division strategy supports a quantity-based representation of fractions as a specified number of parts of a whole. Quantity-based representations provide a unified representation that can be used when solving a large variety of rational number problems, such as modeling, estimating, comparing, and doing arithmetic with fractions. Thus, retention of the strategy should be high. In contrast, the invert-and-multiply strategy relies on a visual, position-based representation, and this representation requires different, special-purpose productions to solve a similar variety of rational number problems, and retention of the productions would be relatively low. However, these specialized productions lead to more efficient performance.

Ideally, instruction could bridge from the more meaningful and grounded strategy of picture division to the more abstract and efficient strategy of invert and multiply, while maintaining high retention. Unfortunately, there is no overlap in the problem representations or the productions used by these two strategies, making it difficult to build from one to the next. Because of this limitation, we developed a third strategy, labeled the common denominator strategy, which builds off the picture division strategy and leads to an efficient and general method for dividing fractions. Because this strategy builds on the picture

division strategy, we first discuss suggestions for teaching the picture division strategy and then outline a model of this new strategy.

The cognitive model suggests a careful sequence of lessons for teaching the picture division strategy. Students should first learn to represent fractions as part-whole quantities. Next, students should be taught to use the picture division strategy on problems that rely on the fewest number of productions - dividing a whole number by a unit fraction. After students have learned this minimum set of five productions, they will need help identifying the group size of non-unit fractions and mixed numbers, identifying the number of smaller divisions in bigger divisions if both numbers contain fractions, and converting remainders to fractional values when needed.

After students have experience with the picture division strategy, the common denominator strategy can be introduced as a more general and efficient strategy. Initially, the common denominator strategy can be tightly grounded by the picture division strategy, and then it can be abstracted to a more efficient algorithm. Both the grounded and abstract versions of the common denominator strategy are illustrated in Table 6. The strategy has five main sub-goals: 1) identify the initial values, 2) find the total number of divisions in the starting amount (which may involve finding a common denominator for the dividend and divisor), 3) identify the size of each group (with this common denominator), 4) divide the total number of division by the group size, 5) simplify the answer. After identifying the initial values, students must figure out the total number of divisions in the starting amount, which is analogous to marking the divisions and counting the total number of divisions. To identify the group size, students must make sure the divisor is a fraction that has the same number of parts-per-whole (denominator) as the dividend. Next, the number of groups is found by dividing the total number of divisions by the group size (i.e. dividing the two numerators), which is analogous to marking the groups on the picture and counting the number of groups. This leads to an answer in appropriate fractional form, although the answer may need to be converted from an improper fraction to a mixed numbers. After linking this strategy to the picture division strategy, a more formal, symbol-based strategy can be abstracted, which relies on converting whole and mixed numbers to fractions and finding fractions with a common denominator and then dividing the numerators and denominators. This strategy retains the quantity-based representations of the picture division strategy while being more efficient and general than this strategy. We have used these analyses to design a set of lessons on fraction division that integrate all three strategies, and we are piloting these lessons with sixth grade students who have no prior experience with fraction division.

Table 6: Example of common denominator strategy for solving 1 1/2 ÷ 3/4

Grounded approach:
Because 1/2 = 2/4, and 1 whole = 4 fourths, 1 1/2 = 6/4.
now have: 6/4 ÷ 3/4
In 6/4, there are 2 groups of 3/4, so the answer is 2.
Abstract approach:
Equivalent fractions: 1/2 = 2/4
Mixed to fraction: have 1 2/4: 1 * 4 = 4; 4 + 2 = 6, so 6/4
Now have 6/4 ÷ 3/4
6 ÷ 3 = 2; 4 ÷ 4 = 1
Answer is 2/1, and because any number divided by 1 is that number, the answer is 2.

In summary, cognitive modeling is a promising tool for evaluating alternative strategies and techniques that can be leveraged in the development of better curriculum material and instructional approaches.

Acknowledgments

This work was supported by NIMH/NRSA training grant 5 T32 MH19983-02 and by a grant from Carnegie Learning. We would like to thank Jay Raspat and his students at North Hills Junior High for participating in the empirical study and Juan Casares, Aaron Powers and Willie Wheeler for their help with the cognitive model for the picture division strategy.

References

Anderson, J.R. (1993). Rules of the mind. Hillsdale, N.J.: Erlbaum.

Anderson, J.R., Corbett, A., Koedinger, K.R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. Journal of the Learning Sciences, 4, 167-207.

Craik, F.I.M., & Lockhart, R.S. (1972). Levels of processing. A framework for memory research. Journal of Verbal Learning and Verbal Behavior, 11, 671-684.

Jakwerth, P. (1999) TIMSS Performance Assessment Results: United States. Studies in Educational Evaluation, 25, 277-281.

Lesh, R. & Landau, M. (Eds.) (1983). Acquisition of mathematical concepts and processes. New York: Academic Press.

Lindquist, M.M. (1989). Results from the Fourth Mathematics Assessment of the National Assessment of Educational Progress. Reston, VA: NCTM.

National Council of Teachers of Mathematics. (2000). Principles and standards for school mathematics. Reston, Va: NCTM.

Paivio, A. (1971). Imagery and verbal process. New York: Holt, Rinehart & Winston.

For Better or Worse: Modelling Effects of Semantic Ambiguity

Jennifer Rodd (jrodd@csl.psychol.cam.ac.uk)
Centre for Speech and Language, Department of Experimental Psychology
Cambridge University
Cambridge, UK
Gareth Gaskell (g.gaskell@psych.york.ac.uk)
Department of Psychology
University of York, York, UK
William Marslen-Wilson (william.marslen-wilson@mrc-cbu.cam.ac.uk)
MRC Cognition and Brain Sciences Unit
15 Chaucer Road, Cambridge, UK

Abstract

Several studies have reported an advantage in lexical decision for words with multiple meanings. More recently, Rodd, Gaskell, and Marslen-Wilson (in press) have reported a more complex pattern of ambiguity effects. While there is a processing advantage for words that have many highly related word senses (e.g., *twist*), there is a disadvantage for words that have more than one meaning (e.g., *bark*). Here we show that these two apparently opposite effects of ambiguity can both emerge from the competition to activate a coherent semantic representation in an attractor network. Ambiguity between unrelated meanings delays recognition because of interference between the two possible stable patterns of semantic activation, that correspond to separate attractor basins. In contrast, the patterns of semantic activation that correspond to different senses of the same word meaning all lie within a single attractor basin, and the semantic flexibility associated with these words results in a widening of the attractor basin, thus produces a processing advantage relative to unambiguous words.

The Ambiguity Disadvantage and Sense Benefit

Models of word recognition often make the simplifying assumption that each word in the language has a single, well-defined meaning. However, many words refer to more than one concept. For example, *bark* can refer either to a part of a tree, or to the sound made by a dog. Other words, such as *twist*, have a range of systematically related dictionary definitions including to *make into a coil or spiral*, *to operate by turning*, *to alter the shape of*, *to misconstrue the meaning of*, *to wrench or sprain*, and *to squirm or writhe*. To understand such words, we select the appropriate interpretation, normally on the basis of the context in which the word occurs. In this paper we review the literature on how semantic ambiguity affects the recognition of single words, and report a series of network simulations that examine the implications of these results for models of word recognition

Several studies in the literature report faster lexical decision times for ambiguous words, compared with unambiguous words (Azuma & Van Orden, 1997; Borowsky & Masson, 1996; Millis & Button, 1989). There have been various explanations for why it might be easier to recognise words with multiple meanings. Typically it is assumed that ambiguous words benefit from having more than one competitor in the race for recognition. More recently, this view that there is a simple advantage for semantic ambiguity has been challenged. Rodd et al. (in press) argue that a distinction should be made between the accidental ambiguity of words like *bark* which, by chance, have two unrelated meanings, and the systematic ambiguity of words that have multiple senses. For example, although there are important differences between what it means to *twist an ankle* compared with to *twist the truth*, these different senses of the word *twist* are closely related to each other, both etymologically and semantically. This relationship is quite unlike the ambiguity for a word like *bark*.

All standard dictionaries respect this distinction between word meanings and word senses; lexicographers routinely decide whether different usages of the same spelling should correspond to different lexical entries or different senses within a single entry. However, although this distinction appears easy to formulate, people will sometimes disagree about whether two usages of a word are sufficiently related that they should be taken as senses of a single meaning rather than different meanings. However, even if there is not always clear distinction between these two different types of ambiguity, it is important to remember that words that are described as ambiguous can vary on a continuum between these two extremes.

Rodd et al. (in press) support the psychological importance of this distinction in a set of lexical decision experiments which show that while multiple related word senses do produce a processing advantage, multiple unrelated meanings delay recognition. Here we report a series of simulations which investigate whether these two apparently opposite effects of ambiguity can both emerge from the competition to produce a coherent distributed semantic representation within an attractor network.

Semantic Competition Models of the Ambiguity Advantage

Joordens and Besner (1994) and Borowsky and Masson (1996) have tried to model effects of ambiguity using a two-layer Hopfield network (Hopfield, 1982) to learn the mapping between orthography and semantics. The models show an advantage for words that are ambiguous between unrelated meanings. The authors argue that this advantage arises because, when the orthography of a word is presented to the network, the initial state of the semantic units is randomly determined. The network must move from this state to a valid finishing state corresponding to the meaning of the word. For ambiguous words there are multiple valid finishing state, and on average, the initial state of the network will be closer to one of these states than for an unambiguous word, where there is only one valid finishing state. However, as discussed above, it is now apparent that ambiguity between unrelated meanings produces an disadvantage, so that there is a discrepancy between the data and the behaviour of these models.

One limitation of these models is that their performance on

the task was surprisingly poor. Joordens and Besner (1994) report an error rate of 74%. These errors often result from the network settling into blend states, which are a mixture of the word's meanings. Gaskell and Marslen-Wilson (1999) have shown that blends between unrelated semantic representations can be relatively meaningless, and may be closer to a different word in the lexicon than to either of the components of the blend. In the Borowsky and Masson (1996) study, these blend states are not considered to be errors; the authors argue that to perform lexical decision it is not necessary to resolve the ambiguity successfully in order for there to be sufficient familiarity to make a successful lexical decision. Although this approach may be appropriate for modelling the specific task of lexical decision, this would severely limit the model in being extended to be a more general model of word recognition. It is the case that, given an ambiguous word in isolation, we are able to retrieve one of its meanings. In contrast, the model would predict that without a contextual bias to direct us to one of the meanings we would get stuck in a blend state that may be quite unlike either of the meanings.

It is possible that the observed ambiguity advantage may be an artefact of this tendency to settle into blend states. Indeed, Joordens and Besner (1994) report that as the size of their network is increased, and performance improves, the ambiguity advantage is eliminated. However, even in these larger networks, the problem of blend states is still present; Joordens and Besner (1994), report a maximum performance level of 48.8% for ambiguous words. In the following simulation, we attempt to improve the overall performance of the network, and investigate how ambiguity affects performance in a network that is able to successfully retrieve the meanings of ambiguous words.

Simulation 1: The Ambiguity Disadvantage

Introduction

While Hopfield networks are known to have limited capacity, the networks discussed above are performing well below the theoretical capacity limit. Hopfield (1982, pg 2556) stated that *"About 0.15 N states can be simultaneously remembered before error in recall is severe"*, where N is the number of units in the network. Therefore the Joordens and Besner (1994) network should be able to learn 45 patterns, and yet the network cannot reliably learn 4 words. This poor performance is because the patterns corresponding to the different meanings of ambiguous words share the orthographic part of their pattern. Hopfield (1982) noted that these networks have a particular difficulty with correlated patterns. Therefore, the simple Hebbian learning rule, which captures the correlational structure of the training set, may not be suitable for learning ambiguous words.

Simulation 1 uses instead the least mean-square error-correcting learning algorithm, which adjusts the weights between units to reduce any error in the activation patterns produced by the current sets of weights. This may therefore alleviate the problem of blend states, as the learning algorithm will change the weights such that these states are not stable.[1]

[1] Kawamoto, Farrar, and Kello (1994) used this algorithm to learn ambiguous words, but they do not report error rates.

Method

Network Architecture The network has 300 units: 100 (orthographic) input units and 200 (semantic) output units. The network is fully connected; each unit is connected to all other units. All units are bipolar; they are either on $[+1]$ or off $[-1]$.

Learning Algorithm All connection strengths were initially set to 0. During each learning trial, the network was presented with a single training pattern, and an error-correcting learning algorithm was used to change the connection strengths. The change in connection strength from a given unit i to a unit j is given by

$$\Delta w_{ij} = x_i(x_j - \sum_k w_{kj}x_k)/3n, \quad i \neq j, \tag{1}$$

where x_i is the activation of unit i and n is the number of units in the network. The learning rate parameter, $1/3$, was selected to provide good performance after a relatively small amount of training.

Training Unambiguous word representations were created by randomly assigning values of $+1$ and -1 to each of the 100 input and 200 output units, such that half the units in each part of the network were assigned $+1$, and the other half -1. For the ambiguous words, a single, randomly generated input pattern was paired with two different output patterns.

In the Joordens and Besner (1994) simulations, the network was trained on only one unambiguous word and one ambiguous word. Here the training set varies between 1 and 16 pairs of ambiguous and unambiguous words, (i.e., 2 to 32 words or 3 to 48 unique semantic patterns). The number of times that each word was presented to the network was varied between 2 and 64 times. The unambiguous and ambiguous words were matched for overall frequency of the orthographic pattern. For each combination of training set size and length of training, the network was trained, and its performance was tested on 200 independent passes; for each pass, a different, independently generated set of training items was used.

Testing Each input pattern was presented to the network, and the output units were all set randomly to $[+1]$ or $[-1]$. Retrieval of the semantic patterns was the result of an asynchronous updating procedure. A unit was selected at random, and its activation was updated by summing the weighted input to that unit. If this input was greater than zero, then the unit was set to $+1$, otherwise the unit was set to -1. This updating continued until a sequence of 1500 updates produced no change in the state of any unit. The network was considered to have settled correctly only if the activation of all its units was correct when it reached a stable state.

Results

For the unambiguous words, the network settled into the correct semantic pattern for over 99.8% of the words, for all the levels of training and set sizes. For the ambiguous words, performance was more variable. The percentage of trials on which the network settled into a correct training pattern for these words is shown in Figure 1 for different amounts of training, and for different sizes of the training set. Importantly, under some conditions, the network was able to settle

correctly into the semantic pattern corresponding to one of the word's two meanings on 98% of the trials. Therefore, the LMS error-correcting algorithm performed substantially better than the Hopfield algorithm on this task.

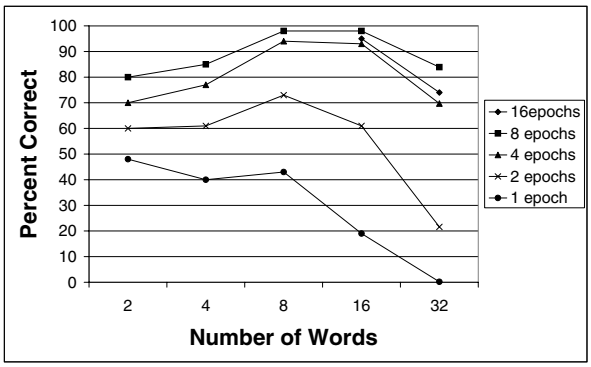


Figure 1: Simulation 1, Performance for Ambiguous Words

Despite this improvement, the ambiguous words were still difficult to learn, compared with the unambiguous words. For the unambiguous words, the network always reached near-perfect performance, having been presented with the training set only once. For the ambiguous words, only when the training set had been presented to the network four times, did performance ever rise above 90%. The number of cycles taken by the network to settle was also generally greater for the ambiguous words than for the unambiguous words. Table 1 shows the difference between the settling times for the two types of words; positive numbers indicate faster settling for the unambiguous words. For the smallest training-set size, the difference between the two types of words was small and variable, but for larger training sets, a consistent ambiguity disadvantage emerges. Crucially, for all the networks where performance on the ambiguous words was greater than 90%, there was a significant ambiguity disadvantage (all significant using the Bonferroni correction for multiple comparisons).

Table 1: Simulation 1, Percentage Benefit in Settling Times for Unambiguous Words

Training	2 Words	4 Words	8 Words	16 Words	32 Words
2	-1	4	38	87	-
4	0	6	23	65	105
8	-1	3	16*	41*	94
16	2	5	11*	25*	64
32	-2	3	12*	32*	57
64	-1	4	11*	32*	49

Notes. * performance on ambiguous words exceeded 90%,

The change in the performance as a function of the size of the training set was somewhat surprising. At all levels of training, the network settled more quickly when it had been trained on fewer patterns. However, the effect of training-set size on error rates for the ambiguous words is more complex (see Figure 1). It is not altogether clear why the network performs so poorly for a very small training sets. It is possible that the increased error produced by the other words in the training set results in the error-correcting learning algorithm operating more effectively; alternatively, the number of spurious stable attractors may increase for small training sets because of the small number of learned attractor basins.

There is an interesting effect of training on performance: initially, training improves performance, in terms of both error rates, and settling times. However, for some training-set sizes, performance reduces if the training set is presented more than 16 times. This suggests that over-learning of the training set produces poor performance for the test items (in which only a subset of the training features are activated).

Discussion

This simulation shows that the introduction of an error-correcting learning algorithm improved performance on the ambiguous words to a level where it is reasonable to investigate the effects of ambiguity on performance. In all conditions where performance exceeded 90%, there was a significant *disadvantage* for the ambiguous words in terms of the number of cycles taken for the network to settle.

Therefore, this simulation suggests that the ambiguity advantage found by Joordens and Besner (1994) is atypical for a network of this type. When performance is improved such that the network reliably settles into a stable semantic representation that corresponds to one of the word's meanings, the interference between the multiple patterns of ambiguous words delays their recognition, relative to unambiguous words. Therefore, a simple semantic competition network of this type can simulate the ambiguity disadvantage seen by Rodd et al. (in press). The question that remains is whether this type of network can also produce the benefit for words with multiple, related word senses.

Simulation 2: Word Senses as Random Noise

Introduction

We have now shown that semantic competition between word meanings delays the settling of the network for ambiguous words, relative to unambiguous words. How then are we to explain the advantage reported by Rodd et al. (in press) for words with multiple senses? One difference between these two forms of ambiguity is the degree of semantic overlap between the alternative semantic patterns. However, although an increase in the similarity of the two meanings of an ambiguous words may reduce the level of semantic competition (and therefore the ambiguity disadvantage), this can only improve performance to the level of the unambiguous words; it cannot produce a benefit.[2]

In this simulation, we explore the hypothesis that the variation in the meanings of words such as *twist* and *flash*, which are listed as having many word senses, should be viewed not in terms of ambiguity, but in terms of flexibility. We assume that the multiple senses of these words are not distinct, but that their meaning is flexible or vague, such that it has a slightly different interpretation in different contexts. In particular, we assume that these words can be represented as having a single base pattern that represents the core meaning of the word. Then, every time this pattern is presented to the

[2]This has been confirmed in a set of simulations identical to Simulation 1 except that the semantic relationship between the meanings of the words was systematically varied (Rodd, 2000).

network, random noise is added to this base pattern, such that each time the network sees the word, it is slightly different from other instances of the word.

Although this idea that words with many senses should be characterized as words whose meanings are flexible about a core meaning does not reflect how these words are listed in dictionaries, there is support for this idea that the classification of the meanings of such words into distinct senses is artificial. For example, Sowa (1993) states that "for polysemous words, different dictionaries usually list different numbers of meanings, with each meaning blurring into the next".

The reason that we might expect this characterization of word senses to produce the processing benefit seen in the human data is that, as we saw in Simulation 1, if an identical pattern is repeatedly presented to the network, it can develop a very deep attractor basin that can be difficult for the network to settle into when it is given only the orthographic input. It is possible that adding a small amount of noise to the network might prevent this over-learning, and might allow the network to develop broader attractor basins, that are easier for the network to enter.

Method

Network Architecture, Learning Algorithm and Processing The architecture and learning algorithm used in this simulation were identical to those used in Simulation 1. However, to reduce the length and variability of the settling times, a different updating procedure was used. Updating now consisted of a series of update sequences in which all the semantic units were updated once in random order.

Training The networks were each trained on 64 words. Half these words were unambiguous, and were presented to the network in exactly the same form on each presentation. The other words had noise added to them; each time these words were presented to the network, a small number of the semantic units were randomly changed from the original base pattern. The number of units that were changed varied from 1 to 5 across different simulations. The number of times that these words were presented to the network was varied from 16 to 128. For each level of training and noise, 100 networks were trained on independently generated sets of patterns.

Results

For the unambiguous words, the network settled correctly in over 99.5% of trials, in all conditions. For the words that had noise added to them, it is less clear what it means for the network to settle correctly; as the level of noise increased, the percentage of trials on which the network settled into the base pattern decreased. However, those trials on which the network did not settle into the base pattern should not all be considered as errors. If the network settles into a pattern that does not differ from the base pattern by more than the amount of noise that was added to the patterns during training, this can be thought of as the network settling into one of the word's senses rather than the core meaning, and should not be considered to be an error. Using this approach, the percentage correct for these words was always above 99.5%

Table 2 shows settling times for the unambiguous words and the words with the added noise, and the differences between these scores. Positive numbers reflect a disadvantage for the noisy patterns. These data show complex interactions between the effects of noise and training, but crucially, while low levels of noise have no stable influence on performance, as the level of noise increases, a reliable disadvantage for noise emerges. This disadvantage for noise is greatest at low levels of training, and increases with the level of noise.

Table 2: Simulation 2, Cycles to Settle for Unambiguous and Noisy Words

Units Changed		Training Presentations 16	32	64	128
1	Unambiguous	603	617	615	588
1	Noisy	611	622	614	593
1	Difference	+8	+5	-1	+5
3	Unambiguous	587	598	564	515
3	Noisy	617	614	583	539
3	Difference	+30	+16	+19	+24
5	Unambiguous	578	573	536	481
5	Noisy	623	609	568	509
5	Difference	+45	+36	+32	+28

Discussion

Contrary to the idea that noise during training might improve performance, the network was slower to settle into those training patterns that had noise added to them during training, compared with the unambiguous patterns. Therefore, this simulation suggests that even if we characterize the ambiguity between multiple senses as being noise about a base pattern, the ambiguity still produces a processing disadvantage. This is, of course, the reverse of the pattern seen in the human data.

In this simulation, the noise that was added to the semantic representations was random; on each training trial, the activation of a given number of units in the semantic pattern was changed. However, this is not a realistic characterization of how the senses of words differ; it is not the case that a new sense of a word can be created from the core meaning of the word by simply changing arbitrary features. Rather, it is that case that these words have sets of possible semantic features which are sometimes, but not always, present. Therefore, rather than modelling word senses as the addition of random noise, it might be better to assume that each word has a range of possible semantic features, but that not all these features are always turned on. For example, the word *twist* may in some contexts not activate the features relating to pain, but it will never arbitrarily gain a feature such as *has legs*.

To model word senses in this way, we need to move away from semantic representations in which half the units are set to $+1$ and the other half set to -1. Instead, for any given semantic representation, most of the units will be turned off, and only a subset will be turned on. Noise is added such that only those features that should be turned on may be turned off, but there is no arbitrary addition of semantic features.

Simulation 3: Sparse Representations

Method

Network Architecture, Learning Algorithm and Processing The architecture used in this simulation was identical to

that used in Simulations 1 and 2. The training patterns used were sparse, such that only 10% of the units were set to $+1$ and the remainder were set to 0. The change in connection strength from a given unit i to a unit j is given by

$$\Delta w_{ij} = 5x_i(x_j - \sum_k w_{kj}x_k)/n, \quad i \neq j \tag{2}$$

Training As in Simulation 2, the network was trained on 64 words; half of the words had noise added to the semantic representations during training, and the other half did not. Again, the number of units that were changed for the noisy words varied from 1 to 5; however noise was added only to units that were set to $+1$ in the base pattern.

Results

Figure 2 shows the performance of the network at different levels of training and noise.[3] At all levels of noise, performance is better for the words that had noise added during training; the network is able to correctly produce the semantic representations for these words at lower levels of training.

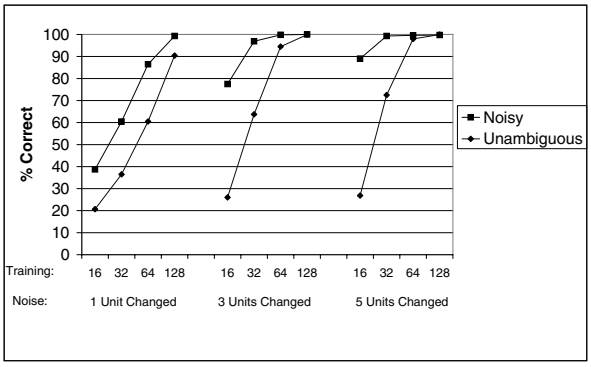


Figure 2: Simulation 3, Error rates

We then looked in detail at the settling behaviour of the network, which was presented with each word 128 times, with a level of noise of 5 units. This network successfully retrieved the meaning in over 99.7% of trials for both types of words. This network settled significantly more quickly for the unambiguous words than for the noisy words; the unambiguous words took on average 407 updates before they were stable, the noisy words took 435 ($t(99) = 8.9, p < .001$). However, a more interesting picture emerges if we look at how the activation of the semantic representations built up over time for this network. Figure 3 shows the total number of semantic units that are switched on at the end of each update of the 200 semantic units. If the network activates 20 units, this corresponds to the activation of a complete semantic pattern. For the noisy words, however, the network tends to activate only a subset of the 20 units; this corresponds to the activation of a sense of the word that does not contain all the possible semantic features for that word.

[3]Unambiguous words are considered to have settled correctly if they settled into the exact training pattern. Noisy words are considered to have settled correctly if they do not differ from their base pattern by more than the amount of noise that was added during training. In a separate analysis, not reported here, this tolerance was also used for the unambiguous patterns; In this analysis, no the error rates differed from those reported here by more than 0.2%.

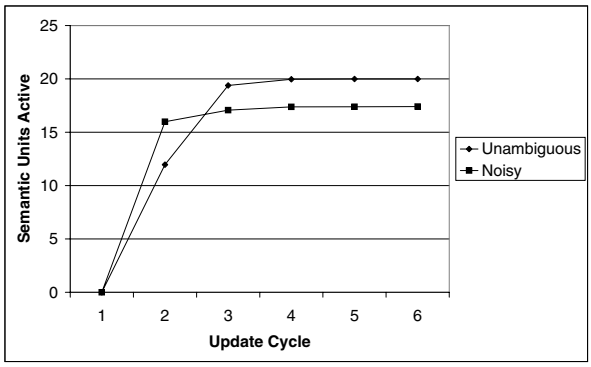


Figure 3: Simulation 3, Activation of Semantic Units

Interestingly, at the end of the first update cycle, the network is significantly more active for the noisy words ($p < .001$). For the unambiguous words, on average 12 units are switched on; for the noisy words, 16 are activated. Therefore, if we assume that lexical decisions are made before the activation of the semantic units has become completely stable, there will be an advantage for the noisy words. It is worth noting that, if this network is presented with a novel word that was not in the training set, the activation of the semantic units very rarely rises above 10. If an activation threshold were set at this level, there would be an advantage for the noisy words.

The later advantage for unambiguous words reflects an assumption built into the training set that the total number of semantic features that are ever activated for words with many senses is equivalent to the total number of features for the unambiguous words. In other words, we have assumed that the individual senses of words with many senses have fewer semantic features than those with only a single sense. This assumption is probably incorrect; it is more likely that words with many senses have a larger set of possible semantic features than words with few senses. It may have been more realistic to assume that the groups of words should be equated on the average number of features that are activated for each individual sense. If this had been the case then the two types of words would settle to the same mean activation level, and the noise advantage would be larger, and extend later in the settling of the network.

Discussion

This simulation shows that if the activation of the semantic features is used as a metric of lexical decision, then there is an advantage for words to which noise is added during training. The advantage is only present early in the settling of the network. This suggests that, as predicted, the noise acts to ensure that the attractor basins are sufficiently wide to allow the activation of the networks to enter the basin quickly. Later in the processing, however, there is a disadvantage for these words; this may be because of competition between multiple stable states (within the large attractor) that correspond to the different senses of the words.[4]

[4]Additional simulations, not reported here, show that the low error rates for ambiguous words and ambiguity disadvantage seen in Simulation 1 is also seen in the rate of activation of semantic units when these sparse representations are used (Rodd, 2000).

Conclusions

The simulations reported here show that networks using the same architecture and learning rule can accommodate the two apparently opposite effects of semantic ambiguity reported by Rodd et al. (in press). While the semantic competition associated with the ambiguity between unrelated meanings delays recognition, the flexibility around the base pattern seen in words with many senses can produce a benefit.

The ambiguity disadvantage shown in Simulation 1 is important because previous simulations of ambiguity effects using networks of this type have shown an ambiguity advantage. We argue that these earlier results were atypical, and relied on using networks that were not able to disambiguate between the different meanings of ambiguous words. Simulations 2 and 3 show that a network of this type can also show a benefit for words whose meanings are flexible between different word senses, but only when their semantic features vary within a limited set of possible features. This limitation fits in with our intuitions about how the semantic representations of words senses vary.

These contrasting effects of ambiguity can best be viewed in terms of the attractor structure of the network. The delay in activating the meaning of an ambiguous word is due to competition between the two stable attractors that correspond to the two different meanings of the word. The initial state of the semantic units produced by the orthographic input will correspond to an unstable blend of the two meanings; the attractor structure of the network will then move the activation of the units away from this blend state towards one of the stable attractors. This disambiguation process takes time, and is responsible for the observed ambiguity disadvantage. In contrast, the different senses of a words all lie within a single attractor basin. Further, the semantic flexibility associated with these words results in a widening of the attractor basin, thus producing a processing advantage relative to unambiguous words. There may, however, be a disadvantage for these words later in processing, due to the existence of multiple stable attractors within the large basin that corresponds to the set of different senses of the word.

In summary, these simulations show that it is possible that the pattern of ambiguity effects reported by Rodd et al. (in press) can be explained in terms of the effects of these two types of ambiguity on the competition to activate a coherent semantic representation within an attractor network. The next stage is to determine whether these explanations are correct.

First, we have assumed that words with many senses should be characterised as words whose meanings are flexible about a core meaning; this assumption must be validated on the basis of detailed analysis of the stimuli used in the experiments. Second, it needs to be confirmed that flexibility is the key property responsible for the sense benefit. As noted by Rodd et al. (in press), words with many senses differ from words with few senses on a range of dimensions, including semantic richness and contextual predictability.

Finally, these simulations investigate two extreme cases of ambiguity; we have compared words with two completely unrelated meanings, with words whose different senses correspond to all the possible combinations of a set of permitted features. This is clearly unrealistic - most words with multiple senses do have some level of structure, and the variation in word senses is often systematic across words. Although the simulations reported here demonstrate important principles about how extreme forms of ambiguity can affect processing, further work needs to be done using more realistic semantic representations. These issues are important if we are to fully understand the implications of ambiguity effects for theories about the representation and access of word meanings.

References

Azuma, T., & Van Orden, G. C. (1997). Why safe is better than fast: The relatedness of a word's meanings affects lexical decision times. *Journal of Memory and Language*, *36*, 484–504.

Borowsky, R., & Masson, M. E. J. (1996). Semantic ambiguity effects in word identification. *Journal of Experimental Psychology: Learning Memory and Cognition*, *22*, 63–85.

Gaskell, M. G., & Marslen-Wilson, W. D. (1999). Ambiguity, competition and blending in speech perception. *Cognitive Science*, *23*, 439–462.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America–Biological Sciences*, *79*(8), 2554–2558.

Joordens, S., & Besner, D. (1994). When banking on meaning is not (yet) money in the bank - explorations in connectionist modeling. *Journal of Experimental Psychology: Learning Memory and Cognition*, *20*, 1051–1062.

Kawamoto, A. H., Farrar, W. T., & Kello, C. T. (1994). When two meanings are better than one: Modeling the ambiguity advantage using a recurrent distributed network. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 1233–1247.

Millis, M. L., & Button, S. B. (1989). The effect of polysemy on lexical decision time: now you see it, now you don't. *Memory & Cognition*, *17*, 141–147.

Rodd, J. M. (2000). *Semantic representation and lexical competition: Evidence from ambiguity.* Unpublished doctoral dissertation, University of Cambridge.

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (in press). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*.

Sowa, J. F. (1993). Lexical structure and conceptual structures. In J. Pustejovsky (Ed.), *Semantics and the lexicon* (pp. 223–262). Dordrecht/Boston/London: Kluwer Academic Publishers.

A Comparative Evaluation of Socratic versus Didactic Tutoring

Carolyn Penstein Rosé (rosecp@pitt.edu)
LRDC, Univ. of Pittsburgh, Pittsburgh PA, 15260, USA

Johanna D. Moore (jmoore@cogsci.ed.ac.uk)
HCRC, Univ. of Edinburgh, Edinburgh EH8 9LW, UK

Kurt VanLehn (vanlehn@pitt.edu)
LRDC, Univ. of Pittsburgh, Pittsburgh PA, 15260, USA

David Allbritton (dallbrit@condor.depaul.edu)
Dept. of Psychology, DePaul Univ., Chicago, IL 60614, USA

Abstract

While the effectiveness of one-on-one human tutoring has been well established, a great deal of controversy surrounds the issue of which features of tutorial dialogue separate effective uses of dialogue in tutoring from those that are less effective. In this paper we present a formal comparison of Socratic versus Didactic style tutoring that argues in favor of the Socratic tutoring style.

Introduction

Comparative studies of student learning have already demonstrated that one-on-one human tutoring is more effective than other modes of instruction. Tutoring raises students' proficiency as measured by pre and post-tests by a minimum of 0.40 standard deviations with peer tutors (Cohen et al., 1982), and to up to 2.0 standard deviations with experienced tutors (Bloom, 1984). One prominent component of effective human tutoring is collaborative dialogue between student and tutor (Fox, 1993; Graesser et al., 1995; Merrill et al., 1992). Nevertheless, a great deal of controversy surrounds the issue of which features of tutorial dialogue distinguish effective uses of dialogue in tutoring from those that are less effective.

In this paper we present the results of a formal evaluation of the relative effectiveness of Socratic versus Didactic tutoring in a simulated problem solving environment in the Basic Electricity and Electronics domain. In this study, the Socratic tutoring style is characterized by an emphasis on eliciting information from students through a directed line of reasoning. Thus, the tutor endeavors as much as possible to avoid giving information away. In contrast, in the Didactic tutoring style, the tutor begins extended interactions with students by presenting the student with an explanation of the material the student is meant to learn in the interaction. After the initial explanation, the tutor leads the student through a directed line of reasoning similar to that used in the Socratic condition, except that the questioning plays more of a role of drawing the student's attention to information that the tutor has already explained, rather than eliciting this information from the student. Since drawing the student's attention to the points already articulated by the tutor requires less from the students, the Didactic interactions tended to be significantly shorter than the Socratic interactions. They contained only 70% as many open ended questions and had more of a lecture like flavor than the Socratic interactions.

In a classroom instructional context, it has been well argued that receptive learning, i.e., by means of lectures, can be just as effective as discovery based learning provided that students have the requisite prior knowledge to learn the presented material meaningfully rather than by rote (Ausubel, 1978). However, this view has not been universally accepted by educational psychologists (Piaget, 1973; Vygotsky, 1978). Furthermore, one key difference between tutoring and classroom learning is that there is much more continuity and regularity in classroom learning. In contrast, tutoring is sporadic and decontextualized. Thus, within a classroom setting the teacher is much more familiar with the students' prior knowledge, and is in fact in a position to ensure that students are prepared to learn the material that is presented each day by arranging lessons to build one upon another. We argue that because this is not the case in a tutoring context, it is more critical to draw out the student's thought process in order to tailor the presentation of material to the students' needs.

Previous studies have argued the effectiveness of Socratic and other similar tutoring approaches. Recent research on student self-explanations supports the view that when students explain their thinking out loud it enhances their learning (Chi et al., 1989, Chi et al., 1994, Renkl, submitted). Students learn more effectively when they are given the opportunity to discover knowledge for themselves (Brown and Kane, 1988; Lovett, 1992; Pressley et al., 1992). Collins and Stevens (1982) report that the best teachers tend to use a Socratic tutoring style. A tutoring system based on the Collins and Stevens model (Wong et al., 1998) has received favorable reviews although it has not yet been subjected to a formal comparative evaluation.

Nevertheless, other studies have argued the effectiveness of Didactic style tutoring. In a recent study in which students read previously solved probability problems (Renkl, submitted), an experimental group that had the option to request further tutor explanation performed better than a control group that did not have that option, with an effect size of .5 sigma. Albacete (1999) found that students who received Didactic conceptual minilessons when they made errors learned more than students who received immediate flag feedback and could

request first a pointing hint, then a correct answer. Similarly, McKendree (1990) found that feedback messages with high content caused more learning than feedback messages with low content. Finally, it is reported in (Graesser et al., 1995) that ordinary human tutors seldom use Socratic tutoring, and yet are quite effective.

The results of our formal comparison presented here demonstrate a trend in favor of Socratic style tutoring over Didactic style tutoring.

Experimental Setup

The context of our work is a web-based course on basic electricity and electronics (BE&E). The system was developed with the VIVIDS authoring tool (Munro, 1994) at the Navy Personnel Research and Development Center in San Diego, CA. The original BE&E tutor was designed as a tool for classroom instruction. As we are interested in one-on-one tutoring, we observed how students interacted with the system and with a human tutor in a Wizard-of-Oz setup. The curriculum used in our experiments and prototype system consists of four lessons and six labs covering basic concepts of current, voltage, resistance, power, making measurements with a multimeter, and doing some simple computations and problem solving. Each lesson consists of three sections of between 10 and 25 pages of instructional text and graphical illustrations displayed in a Netscape window, as in Figure 1. After each section, the student was tested with between 3 and 5 multiple choice progress check questions, which act as a springboard for interaction with the tutor to address deficiencies in the student's understanding. After each lesson, the student was presented with one or two labs designed to test and reinforce the concepts introduced in the lesson. The students completed the labs by interacting with a simulated electronics workbench through a point-and-click interface, as in Figure 2.

The 37 subjects who participated in our data collection experiment were University of Pittsburgh undergraduates with little or no prior study of electricity or electronics. Each subject participated in two sessions, each of which lasted between two and two and a half hours. The tutor was a post-doc working at the Learning Research and Development Center at the University of Pittsburgh with some tutoring experience. While the student interacted with the system, the video signal to the student's monitor was split so that a tutor sitting behind a partition could watch the student's progress. The student and tutor had access to a chat interface that allowed them to type messages to each other. Although students sometimes initiated dialogues themselves, the majority of dialogues occurred when the tutor initiated a dialogue because the student either incorrectly answered an important progress check question or showed evidence of not being able to proceed with a lab.

Seventeen of the students participated in a pilot study, which we used to determine how much material on average that students are able to cover in the allotted time (five hours maximum) and which concepts from the domain come up most frequently in the tutoring dialogues.

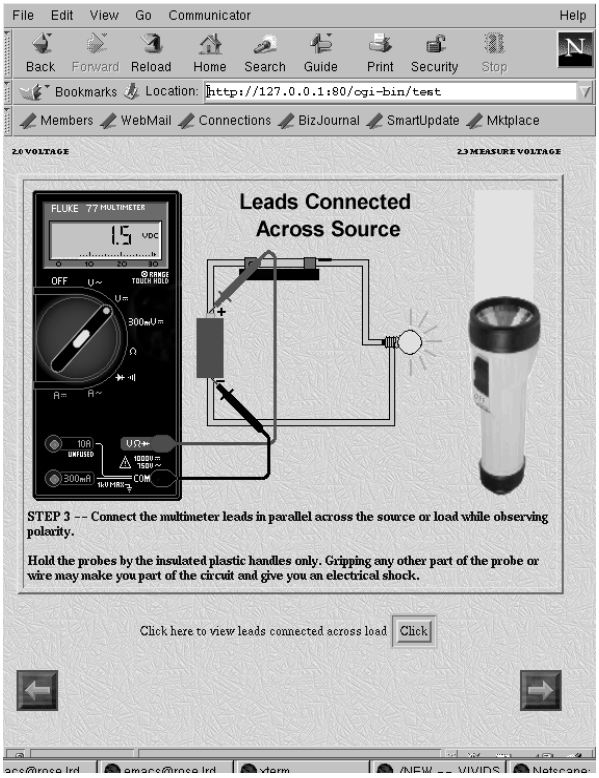


Figure 1: **Netscape Window**

Thus, we designed the pre/post-test for the formal study to focus on these troublesome concepts. We also slightly shortened the lessons by removing some material not essential to learning these concepts (e.g., how to interpret the colored stripes on resistors).

Twenty students took part in the final data collection effort. Each student was randomly assigned to either the Socratic condition or the Didactic condition, described above. Note that the Socratic tutoring style in this study did not typically include heavy use of Socratic irony (i.e., proof by contradiction) as a teaching tool.

Examples of Socratic and Didactic tutoring dialogues from our corpus are displayed in Figures 4 and 5 respectively. Both dialogues occurred during a lab in which the students were expected to find three places in a DC circuit where they could get a non-zero voltage reading. A very typical error students made was to attempt to get a non-zero voltage reading across a closed switch. Both the Socratic and Didactic dialogue examples occurred in response to this error. The important piece of information the tutor wanted to get across to the student in both cases was that it is only possible to get a non-zero voltage reading where there is a difference in charge, and there is no difference in charge across a switch. Notice that in both cases the tutor presented the students with questions to encourage the students to think through the reasoning behind their incorrect action. However, in the Socratic example, the student is doing more of the talking,

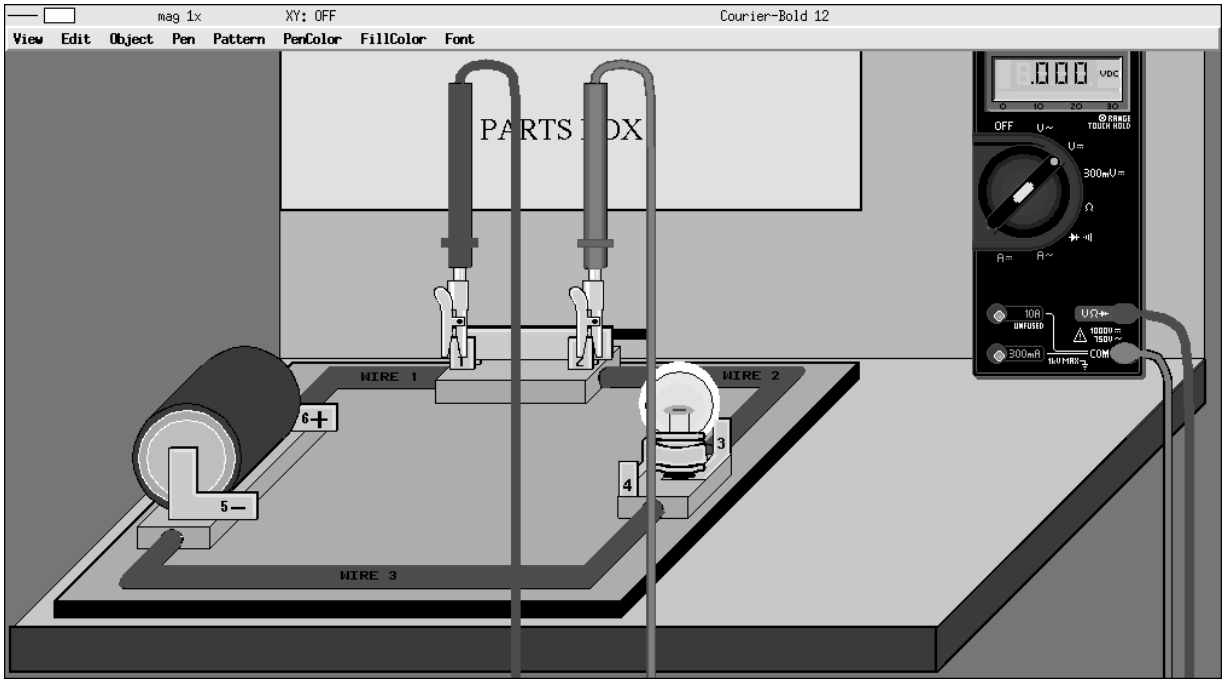


Figure 2: **Simulation Window**

where the Didactic example contains much more lecturing on the tutor's part. In the Didactic example, the tutor explained all that the student needed to know before asking the student any contentful questions. In contrast, in the Socratic example, the tutor asked contentful though questions from the beginning and explained as little as possible. In general, because in the Didactic condition the tutor was able to refer to parts of her own explanation that initiated the dialogue, the Didactic dialogues contained approximately 70% as many open ended questions such as why and how questions. Because in the Socratic condition the students were responsible for articulating as many key concepts as possible themselves, these open ended questions were essential for drawing out the students' reasoning.

Data Analysis

The data collected from the twenty students who participated in the formal study was used in a rule gain analysis to determine the relative effectiveness of the two alternative tutoring strategies. As in (VanLehn et al., 1998), the purpose of our rule gain analysis was to compute a correlation between student learning and the distinctive features of the two alternative tutorial styles. For each student we kept detailed records of their participation in our study as well as other information relevant to evaluating their learning. In particular, we recorded which of the two conditions the student was assigned to, the student's math and verbal SAT scores, the amount of time spent on each lesson and lab, the full text from all tutorial dialogues, and a pre and post-test. The same test was used for the pre-test as for the post-test. Note that students in

```
You can only get a non-zero voltage
reading across a portion of a circuit
where there is a difference in charge
(voltage potential)

There is a difference in voltage
potential across a source

There is a difference in voltage
potential across a load

Sources cause current to move because of
a difference in potential between their
two ends

Loads have a difference in potential
between their two ends because their
resistance causes them to convert
electrical energy into energy of a
different form
```

Figure 3: **Sample Rules for Difference in Voltage Potential**

the Didactic condition spent slightly more time interacting with the tutor and system overall (Socratic students spent an average of 192.68 minutes, SD=37.0; Didactic students spent an average of 198.72 minutes, SD=47.19), however the difference was not significant ($t(18)=.318$, $p=.754$).

For analyzing student learning, we codified the concepts we wanted to teach students into a list of 47 rules. For example, the five rules expressing basic concepts about voltage potential and voltage measurements are

```
Tutor:     So now let's think about the switch.  Is it a source or a load?

Student:   source

Tutor:     Why do you say that?

Student:   when it is turned off, no energy flows

Tutor:     OK, that is true.  But does it actually cause electricity to flow?
           Or is it merely letting what current was produced by the source
           flow through it to the rest of the circuit?

Student:   it does not create energy

Tutor:     Good.  So it is not a source.  Would you say it is a load then?

Student:   no

Tutor:     Good, it is neither a source nor a load.  So does it have a
           difference in charge between its two ends?

Student:   no

Tutor:     Very good.  So then can you measure voltage across a switch?

Student:   no

Tutor:     Right.  So how can you modify the way you have the leads hooked up
           now so that there will be a difference in charge between where the
           two leads are hooked up?

Student:   since i know there is a third place to hook them up, is it possible
           to connect the one to the negative side of the battery and connect
           the other to the positive of the light bulb?

Tutor:     Yes, that would be perfect.  Go ahead and do that.
```

Figure 4: Sample Socratic Dialogue

found in Figure 3. The pre/post test consists of 40 questions that require the students to use the rules either by applying a rule in order to solve a problem, directly articulating a rule, drawing a diagram to illustrate a rule, answering a question about a rule, or recognizing a diagram illustrating a rule. We coded each question in our pre/post test with a list of rules that applied to solving that problem and how those rules applied. For each rule we then developed a formula to assess the student's level of mastery of that rule based on which of the questions where the rule applied the student answered correctly. The formulas designed for computing the mastery score for each rule were based on the assumption that different ways rules apply to questions give different amounts of evidence about how well students have mastered the corresponding rule. For example, directly articulating a rule gives more evidence of student knowledge than answering a multiple choice question. Solving a problem by using a rule gives an even stronger indication. A total mastery score for each student was computed for both the pre and post-test by summing the mastery scores for the 47 individual rules. We also computed a gain score for each student by counting the number of rules for which each student demonstrated a higher mastery score on the post-test as compared with that on the pre-test. For each rule we noted how many students in each condition achieved a higher mastery score for that rule on the post-test as compared to the pre-test.

An ANCOVA with condition as the independent variable, pre-test score as the covariate, and post-test score as the dependent variable confirmed a trend for students in the Socratic condition to learn more (pre-test mean = 10.41, pre-test SD = 7.5, post-test mean = 27.54, post-test SD = 7.28) than students in the Didactic condition (pre-test mean = 14.29, pre-test SD = 53.12, post-test mean = 25.5, post-test SD = 6.31), $F(1, 18) = 3.13$; $p < .1$. Interestingly, despite the fact that the students in the Socratic condition had a lower average pre-test score than the students in the Didactic condition, they achieved a higher average post-test score. The effect size (mean Socratic gain score - mean Didactic gain score / SD Didactic gain score) was 1 sigma. Additionally, for each rule we computed a chi-square to determine whether the number of students who demonstrated learning on that rule in the Socratic condition was significantly higher than the number of students who demonstrated learning on that rule in the Didactic condition. The difference was only significant ($p < .05$) for two rules although for every rule more students in the Socratic condition than in the Didactic condition demonstrated learning. The probability that more students in the Socratic condition would demon-

```
Tutor:      Do you remember in the lesson that it said that every point on a
            conductor is electrically the same?

Student:    yes.

Tutor:      Good.  That means that there is no difference in potential energy
            between one point on the conductor and another point.  So, if both
            leads are attached to the same conductor, there is no difference
            in potential (in other words, no difference in charge) and thus no
            force to measure.  What you need is to have there be a difference
            in potential between where the red lead is attached and where the
            black lead is attached.  This is achieved whenever voltage is
            "created" as in a battery or "used up" as in a light bulb.  So,
            where do you suppose you could attach the leads now to achieve
            that?

Student:    to the light bulb, it seems to be my only source

Tutor:      Right.
```

Figure 5: Sample Didactic Dialogue

strate learning for all 47 rules by chance is very small, specifically $.5^{47}$. We found during our analysis, however, that in spite of having randomly assigned students to conditions, the average SAT score for students in the Socratic condition (1161, SD = 208) was marginally higher than that for students in the Didactic condition (961, SD = 192), F (1, 15) = 4.27, $p < .06$. Finding even a marginal trend with only 20 subjects suggests that the effect may be real even if the statistical power is not sufficient.

A re-analysis in which gain score (post test score minus pretest score) was the dependent variable and SAT score was a covariate was conducted for the 17 participants for whom SAT scores were available. The effect of condition was not significant in this analysis, F (1, 14) = 1.64, p ¿ .20, although the loss of power resulting from the exclusion of three participants from this analysis may have contributed to the lack of statistical significance. Thus, although the trend for greater gains in the Socratic condition was still evident after controlling for SAT scores, we can not conclusively rule out the influence of this possible confound on the results.

Next we checked for an aptitude-treatment interaction by dividing the students into two subsets, with students having above the mean SAT scores in one subset and those with below the mean SAT scores in the other subset. When the ANCOVA was performed on each subset separately, a trend was demonstrated for students in the Socratic condition to perform better within each subset. For above average SAT students, F (1,7) = 1.77, $p < .23$. For below average SAT students, F (1,4) = 5.62, $p < .1$. This seems to be a result of the fact that the below average SAT students tended to have uniformly low pre-test scores and variable post-test scores whereas the above the average SAT students had variable pre-test scores and uniformly high post-test scores.

Discussion and Current Directions

In this paper we present a study in which we explore the relative effectiveness of two alternative tutoring styles, which we have referred to as Socratic versus Didactic. The purpose of our study was to explore alternative methods of encouraging knowledge construction via tutorial dialogue in order to determine how to use dialogue most effectively to this end. The results of our rule gain analysis demonstrate that students in the Socratic condition learned more effectively than students in the Didactic condition, although more data collection is necessary in order to verify the level of statistical significance.

Based on our findings, we are building a dialogue-enhanced version of the original BE&E tutoring system. A small prototype dialogue based version has already been built covering the portion of our curriculum concerned with teaching about measuring voltage in DC circuits (Rosé et al., 1999). Note that in the example Socratic dialogue in Figure 4, the tutor affirmed what was correct in the student's response, that when a circuit is turned off no energy flows, and then addressed specifically what was lacking in the student's explanation. The same ability to evaluate the content of student explanations is required for the tutor to determine when it is no longer beneficial to continue prompting a student with leading questions. This type of sensitivity in tutor response is only possible in an intelligent tutoring system when the system can understand what the student says. Thus, a major focus of our work has been on robust natural language understanding (Rosé, 2000). Our robust core understanding component for English is currently being integrated into three different tutoring systems and is available for use on other projects[1].

[1] Parties interested in obtaining the Atlas core understanding component should contact Carolyn Rosé at `rosecp@pitt.edu`.

References

Albacete, P. L. (1999). *An Intelligent Tutoring System for Teaching Fundamental Physics Concepts*. PhD thesis, University of Pittsburgh, Pittsburgh, PA.

Ausubel, D. (1978), *Educational Psychology: A Cognitive View*, Holt, Rinehart and Winston, Inc.

Bloom, B. S. (1984). The 2 Sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13:4–16.

Brown, A. L. and Kane, M. J. (1988). Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology*, 20:493–523.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., and Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2):145–182.

Chi, M. T. H., de Leeuw, N., Chiu, M. H., and LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3):439–477.

Cohen, P. A., Kulik, J. A., and Kulik, C. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, 19:237–248.

Collins, A. and Stevens, A. (1982). Goals and methods for inquiry teachers. In Glaser, R., editor, *Advances in Instructional Psychology, Vol. 2*. NJ: Lawrence Erlbaum Associates, Hillsdale.

Fox, B. A. (1993). *The Human Tutorial Dialogue Project: Issues in the design of instructional systems.* Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Graesser, A. C., Person, N. K., and Magliano, J. P. (1995). Collaborative dialogue patterns in Naturalistic One-to-One Tutoring. *Applied Cognitive Psychology*, 9:495–522.

Lovett, M. C. (1992). Learning by problem solving versus by examples: The benefits of generating and receiving information. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*. NJ:Erlbaum.

McKendree, J. (1990). Effective feedback content for tutoring complex skills. *Human-Computer Interaction*, 5:381–413.

Merrill, D. C., Reiser, B. J., and Landes, S. (1992). Human tutoring: Pedagogical strategies and learning outcomes. Paper presented at the annual meeting of the American Educational Research Association.

Munro, A. (1994). Authoring interactive graphical models. In de Jong, T., Towne, D. M., and Spada, H., editors, *The Use of Computer Models for Explication, Analysis and Experiential Learning*. Springer Verlag.

Piaget, J. (1973). To understand is to invent. New York: Grossman.Resnick, L. B. (1989). Developing mathematical knowledge. AmericanPsychologist, 44, 162-169.

Pressley, M., Wood, E., Woloshyn, V. E., Martin, V., King, A., and Menke, D. (1992). Encouraging mindful use of prior knowledge: Attempting to construct explanatory answers facilitates learning. *Educational Psychologist*, 27:91–109.

Reiser, B. J., Copen, W. A., Ranney, M., Hamid, A., and Kimberg, D. Y. (in press). Cognitive and motivational consequences of tutoring and discovery learning. In *Cognition and Instruction*.

Renkl, A. (submitted). Worked-out examples: Instructional explanations support learning by self-explanations.

Rosé, C. P. (2000). A framework for robust semantic interpretation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*.

Rosé, C. P., Di Eugenio, B., and Moore, J. D. (1999). A dialogue based tutoring system for basic electricity and electronics. In *Proceedings of the Ninth World Conference on Artificial Intelligence in Education*.

VanLehn, K., Siler, S., and Baggett, W. (1998). What makes a tutorial event effective. In *Proceedings of the Twenty-first Annual Conference of the Cognitive Science Society*.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., and Baggett, W. B. (in press). Human tutoring: Why do only some events cause learning? cognition and instruction.

Vygotsky, L. S. (1978). Mind in society: The development of higherpsychological processes (M. Cole, V. John-Steiner, S. Scribner, & ESouberman, Eds.). Cambridge, MA: Harvard University Press.

Wong, L. H., Quek, C., and Looi, C. K. (1998). TAP-2: A framework for an inquiry dialogue-based tutoring system. *International Journal of Artificial Intelligence in Education*, 9.

Acknowledgments

This research is supported by the Office of Naval Research, Cognitive and Neural Sciences Division (Grants N00014-91-J-1694 and N00014-93-I-0812) and NSF Grant 9720359 to CIRCLE, a center for research on intelligent tutoring.

Mental Models and the Meaning of Connectives: A Study on Children, Adolescents and Adults

Katiuscia Sacco (sacco@psych.unito.it)
Monica Bucciarelli (monica@psych.unito.it)
Mauro Adenzato (adenzato@psych.unito.it)

Centro di Scienza Cognitiva
Universita' di Torino
via Lagrange, 3 10123 Torino, Italy

Abstract

We present a study on the ability to comprehend conjunction, exclusive disjunction, bi-conditional and conditional. Mental model theory predicts differences in difficulty in dealing with such connectives, and it also predicts that it is easier to envisage situations that comply with an assertion than situations which do not comply. We carried out an experiment on children, adolescents and adults to validate these predictions within a developmental perspective. Participants had to judge some states of affairs as complying or not complying with sentences involving connectives. A further aim of the experiment was to test the power of the theory to account for connectives' comprehension within a pragmatic context. Thus, while half of the participants dealt with an abstract version of the task, the other half coped with a pragmatic version where the sentences were uttered by a character known as sincere in the complying condition and as a liar in the not complying condition. The results of the experiment show that difficulty in comprehension of the different connectives depends on the number of models they require. Also, the results show that it is easier to envisage situations complying with the meaning of a connective than situations which do not comply. The results hold for all groups of participants in both versions of the task. We conclude that mental model theory offers a plausible account of connectives' comprehension, which holds also within the investigated pragmatic context.

1. Introduction

Experimental data show that connectives vary in difficulty of comprehension. Conjunction *and* is handled by 2 years old children (Bloom, Lahey, Hood, Lifter & Fiess, 1980); disjunction *or* is understood after 4 years (Johansson & Sjolin, 1975); bi-conditional *only if then* emerges from 8 years of age but it is not fully mastered until 11;6 years (Staudenmayer & Bourne, 1977); conditional *if then* remains difficult even for 14 year old children, although around 5-6 years there is a clear improvement (Amidon, 1976; Staudenmayer & Bourne, 1977).

The connective interpretation varies from children to adolescents to adults. Conjunction is early understood as implying the co-occurrence of its constituents, even if children sometimes treat it as a disjunction (Johnson-Laird & Barres, 1994). Disjunction is commonly interpreted exclusively, namely as if it would imply a choice between the co-ordinated members and they should not be taken in combination; this kind of interpretation seems the favourite at every age (Staudenmayer & Bourne, 1977). To comprehend conditional relations, individuals have to grasp the possibility of relations between properties that are absent, but implied. The first achievement of this kind is the bi-conditional interpretation. Conditional interpretation is hardly caught even by adolescents and adults, who often interpret it as it would imply its converse, i.e. they usually give a bi-conditional interpretation (Taplin, 1971).

The meaning of the connectives has been mainly a concern of the theories on propositional reasoning, viz. the ability to reason with propositions and connectives. Some of these theories claim that the meaning of the connectives is conveyed by the truth-values they receive in a truth table system. For instance, Piaget and Inhelder (Inhelder & Piaget, 1958; Piaget, 1953) claim that people can construct true and false contingencies of propositions because they posses a mental logic: in their view, some truth functions and a set of transformations would develop by the early teens, so that children would grasp the meaning of the connectives.

Other theorists have proposed that the meaning of the connectives is grasped through natural deduction systems, where rules are claimed to have more psychological plausibility than standard logic (Braine, 1978; 1990; 1998, Braine & Rumain, 1983, Rips, 1990). In their view, the evaluation of contingencies complying with propositions would depend on the internal structure of the proposition itself: each connective would define which inferential rules can be applied and, therefore, how reasoners can envisage the correct contingencies.

In a radically alternative view, Pollard (1981) and Griggs and Cox (1982) claim that the understanding of the meaning of connectives depends on the reasoners' previous experience. In particular, the specific experiences encoded in memory would provide a set of domain dependent rules that reasoners can use in the current situation by analogy.

Cheng, Holyoak and colleagues analyze just the conditional connective and postulate the existence of abstract knowledge structures such as causation, obligation and permission (Cheng & Holyoak, 1985; Cheng, Holyoak, Nisbett & Oliver, 1986). They argue that the meaning of a conditional emerges from the concrete context within which it is introduced. For instance, a permission context would induce the reasoner to think about the possibility in which an action is done provided the precondition is satisfied.

All of the mentioned theories offer an explanation of how people represent the meaning of the connectives, but none of them gives a systematical account of their difference in terms of difficulty of comprehension.

Mental model theory offers such an account. Our investigation into the mental representations of the meaning of the connectives follows the tenets of this theory.

2. Mental Models for Connectives

Mental model theory (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) claims that the meaning of the connectives is conveyed by mental models. A model is an analogical representation: it does reproduce the structure of the states of affairs perceived or described. For example, the theory claims that the meaning of an indicative conditional such as:

If you eat too many cakes, then you put on weight

is represented by the model representation:

[too many cakes] put on weight
...

Each row in the representation denotes a model of a separate possibility. The first model is explicit and satisfies the antecedent and the consequent, and the second model (dots) is implicit: it allows for the case in which the antecedent is not satisfied.

The construction of models is guided by the Principle of truth and the attempt to maintain as much information implicit as possible. According to the Principle of truth, each model represents only what is true in a particular possibility. Hence, the first model represents the possibility in which the antecedent (and then the consequent) is true. Reasoners do not represent the possibilities in which the antecedent is false. The theory postulates that reasoners make "mental footnotes" to keep track of this information, but that these footnotes are soon likely to be forgotten. To indicate these footnotes we use the square brackets and the dots. The square brackets indicate that the antecedent (i.e. too many cakes) has been exhaustively represented in relation to the occurrence of the consequent (i.e. put on weight), i.e. it can not occur in any other model of the conditional (see Johnson-Laird, Byrne & Schaeken, 1992). The dots denote the wholly implicit models, in which the antecedent is false. Therefore, the fully explicit representation of a conditional calls for three models. In our example:

too many cakes	put on weight
not- too many cakes	put on weight
not- too many cakes	not- put on weight

In essence, mental model theory (MMT hereafter) assumes that human reasoners tend not to represent information explicitly. In fact, the more information that has to be represented explicitly, the greater the load on the working memory, and so the initial models of a proposition represent as much information implicit as possible. Implicit information is not immediately available for processing, but it becomes available when it is made explicit (see e.g. Bara, Bucciarelli, Johnson-Laird & Lombardo, 1994).

The fully explicit models required by each connective and the implicit models which people tend to construct are in Table 1. In fact, constraints of working memory prevents people to imagine all the possible models of a connective, and because truth appears to be highly salient, people represent first what is true.

Table1: Mental models representing the meaning of connectives (fully explicit models).

Connectives	True instances		False instances	
	Implicit models	Explicit models		Explicit models
p *and* q	p q	p q		p ¬q ¬p q ¬p ¬q
p *or else* q	p q	p ¬q ¬p q		p q ¬p ¬q
only if p *then* q	p q …	p q ¬p ¬q		p ¬q ¬p q
if p *then* q	p q …	p q ¬p q ¬p ¬q		p ¬q

The models representing the false instances of a proposition (see Table 1) would be fleshed out afterwards, only if they are needed to make the deduction. Thus, the theory predicts that the mental representation of cases complying with the meaning of a connective (i.e. the true instances) is easier than that of not complying cases (i.e. the false instances). The prediction is confirmed by Barres and Johnson-Laird (1997). They carried out a study where the participants were asked to list the true and the false instances given an assertion, and they found that representing the false instances is more difficult than representing the true ones. Thus, they claim that there is no a direct way to imagine what is false and errors are likely to occur when listing the false instances. For example, given the assertion "A or B, or both", most of their subjects perform correctly, and list the following instances in which the assertion is true:

A

B

A B

Then, in order to infer the false instances, most subjects negate the true ones and list what follows:

not A

not B

not A not B

while we know the only false instance is:

not A not B

The aim of our experiment is to validate the following predictions within a developmental perspective. First, the difficulty of comprehension of the different connectives depends on the number of models they require. Second, in line with the study by Barres and Johnson-Laird (1997), to envisage the false instances of a connective is harder than to envisage the true instances. Also, we derive a corollary prediction from MMT and from the fact that working memory abilities, such as encoding abilities and the time information which can be maintained in memory (Cowan, 1997; Towse, Hitch & Hutton, 1998), are good predictors of the performance of subjects belonging to different age groups. Thus, the ability to deal with the not complying conditions should increase with age; such ability requires keeping in memory the true instances of a connective while deriving the false ones.

We tested these predictions using two different protocols: one in which the connectives are presented within an abstract context and the other in which they occur in a pragmatic context, where they are uttered by a character describing a certain state of affairs. We expect the evaluation of instances complying with the utterance proffered by a sincere character to be easier than the evaluation of instances not complying with the utterance proffered by a liar. This prediction parallels the prediction concerning complying versus not complying conditions in the abstract version of the task. In particular, granted that 7 year olds do possess the ability to think in term of lies, the requested abilities in the two contexts might be the same. Thus, we should detect an analogy between evaluating true instances of a sentence and understanding a person who is telling the truth, and between evaluating false instances of a sentence and understanding a person who is lying. In previous studies, MMT has been proved to account for the ability to comprehend connectives within different contexts, calling for the same basic principles (see, e.g. Bara, Bucciarelli & Lombardo, 2001). As people in everyday life have to deal with *utterances* rather than with abstract *sentences*, MMT for the meaning of connectives has to hold within pragmatic contexts as well as in abstract contexts.

To sum up, our aim is a validation of the following predictions:

i. The difficulty of comprehension of the meaning of the connectives depends on the number of models they require. Thus, we expect to observe the following trend of difficulty, from the easiest to the most difficult connective: conjunction (one model), exclusive disjunction and bi-conditional (two models), conditional (three models).

ii. The evaluation of cases not complying with a connective is more difficult than the evaluation of complying cases (from the Principle of truth). It requires first to represent the states of affairs consistent with the connective, then to negate them. Also, the ability to evaluate instances not complying with a connective would improve with age.

We expect these predictions to hold both in the abstract and in the pragmatic version of the task.

3. Experiment

Method

Participants. We tested a sample of 180 subjects, 60 in each of the following age groups: children from 7 to 7;9 years old, adolescents from 14 to 14;9 years old, adults from 21 to 24;9 years old. They were students from primary schools, high-schools and university residences, who took part in the experiment voluntarily. There was a balanced proportion of males and females in each group of participants.

Design. We devised two protocols: the Abstract Protocol and the Pragmatic Protocol. In the Abstract Protocol propositions were presented within an arbitrary context, whereas in the Pragmatic Protocol the context was provided by a character proffering the utterance. The participants were randomly assigned to one of the two protocols. Thus, half of the participants of each age group dealt with the Abstract Protocol and half with the Pragmatic Protocol.

In either protocol, subjects had to deal with two conditions: «complying» and «not complying». As for the Abstract Protocol, propositions in the complying condition were said to be true, while propositions in the not complying condition were said to be false. As for the Pragmatic Protocol, propositions in the complying condition were uttered by a character said to be sincere, while propositions in the not complying condition were uttered by a character said to be a liar. Each subject dealt with 8 propositions, 4 in the complying condition and 4 in the not complying condition. The order of presentation of propositions within each condition was determined at random.

In either protocol, after reading the proposition, the experimenter showed the subjects a set of 4 cards, each representing a possible state of affairs. For each card the experimenter asked the subject if it satisfied the proposition or not. The cards were presented in a random order.

Materials. In the Abstract Protocol we used the following materials:

- 8 sheets of paper; on each of them it was written a proposition containing one of the following connectives: *and*, *or*, *only if-then*, *if-then*. Each connective occurred in two propositions, but with different content;
- 8 series of cards: each series consisted of 4 cards. Given a proposition «A *connective* B» (for example, «There are an aeroplane *and* a car»), the four cards represented A and B together (aeroplane and car), A alone (aeroplane), B alone (car), and CD, two things different from the ones mentioned in the proposition (for example, train and boat). Four series of cards were used in the complying condition and four series in the not complying condition.

In the Pragmatic Protocol we used the following materials:
- the puppets Minnie and Lucy;
- 8 sheets of paper; on each of them it was written an utterance proffered by a character. Each utterance contained one of the following connectives: *and*, *or*, *only if-then*, *if-then*. Each connective occurred in two utterances, but with different content;
- 8 series of cards: each series consisted of 4 cards, as in the Abstract Protocol. Four series of cards were used in the complying condition and four series in the not complying condition.

Procedure. The participants were tested individually in a quiet room.

Abstract Protocol: complying condition. The participant was told that he will be presented with some true propositions. Then the experimenter showed the sheet of paper with the first proposition and read it. For example:

> «Either there is a parrot, or there is a fish, but not both». I'll show you some cards: you have to choose those satisfying the proposition.

Then, the experimenter showed one of the four cards (for example, the card representing the fish) and asked:

> Does a fish alone[1] satisfy the proposition?

and waited until the participant has judged the card as satisfying or not the proposition. Then, the experimenter showed another card of the set, questioned the participant and so on with the other cards.

The same procedure was followed with the other three propositions of the complying condition.

Abstract Protocol: not complying condition. The participant was told that he will be presented with some false propositions. The procedure was the same as that in the Abstract Protocol, complying condition.

Pragmatic Protocol: complying condition. Participant was introduced to a character, Minnie, said to be sincere and they are told that Minnie will have proffered utterances about some cards she owns. The experimenter began with the first proposition. For example:

> Minnie says «On each of my cards, only if there is a candle, then there is a book». Remember that Minnie always says the truth. I'll show you some cards, and you have to tell me which cards belong to Minnie.

Then, the experimenter showed one of the four cards (for example, the card representing the book) and asked:

> Can a book alone belong to Minnie?

and waited until the participant has judged the card as belong or not to the sincere character. Then, the experimenter showed another card of the set, questioned the participant and so on with the other cards.

The same procedure was followed with the other three propositions of the complying condition.

Pragmatic Protocol: not complying condition. Participant was introduced to a character, Lucy, said to be a liar and they are told that Lucia will have proffered utterances about some cards she owns. The procedure was the same as that in the Pragmatic Protocol, complying condition.

4. Results

The score was computed assigning one mark for each correct response (the choice of a card which would have to be chosen and the non-choice of a card which would have not to be chosen). So, the maximum score participants could obtain in each trial was 4.

i. Trend of difficulty in comprehension of the different connectives.

The trend in difficulty of comprehension of the different connectives is confirmed in the Abstract Protocol (see Table 2).

Table 2: Mean scores obtained by participants in the Abstract Protocol.

Age groups	Connectives			Mean score
	and	*or/only if-then*	*if-then*	
7-7;9	3.28	2.85	1.97	2.70
14-14;9	3.58	3.26	1.88	2.91
21-24;9	3.87	3.33	2.10	3.10
Overall	3.58	3.15	1.98	2.90

Participants find it easier to comprehend the meaning of «*and*» than the meaning of «*or/only if-then*»: the difference is statistically significant in each age group (paired T Test; t value ranging from 2.607 to 9.406, p value ranging from <.001 to p=.007). Also, participants find it easier to comprehend the meaning of the connectives «*or/only if-then*» than the meaning of «*if-then*»: again, the difference is statistically significant in each age group (paired T test; t value ranging from 6.520 to 19.746, p value is always <.001).

[1] The cards are intended to correspond to the instances pq, p not q, not p q and not p not q. Thus, as the participant received pq, p, q and rs, we wanted to clarify that an implicit negation is intended.

Also, the results show that the knowledge of the meaning of the different connectives does increase with age (ANOVA one-way; F= 7.593, p<.001).

The same results hold in the Pragmatic Protocol (see Table 3).

Table 3: Mean scores obtained by participants in the Pragmatic Protocol.

Age groups	Connectives			Mean score
	and	*or/only if-then*	*if-then*	
7-7;9	3.12	2.63	2.02	2.59
14-14;9	3.55	3.18	2.12	2.95
21-24;9	3.37	3.22	2.20	2.93
Overall	3.35	3.01	2.11	2.82

Participants find it easier to deal with «*and*» than with «*or/only if-then*», and the difference is statistically significant in each age group (paired T Test; t value ranging from 3.447 to 3.768, p value ranging from <.001 to p=.005). An exception are adults: their performance with the different connectives is in the predicted direction, however, the difference is not statistically significant (paired T Test; t= 1.260, p=.109). Further, the participants find it easier to deal with the connectives «*or/only if-then*» than with «*if-then*», and the difference is statistically significant in each age group (paired T Test; t value ranging from 4.389 to 12.853, p value is always <.001).

The results show that also within the pragmatic context the knowledge of the meaning of the different connectives does increase with age (ANOVA one-way; F= 7.142, p<.001).

ii. The evaluation of cases complying with a connective is easier than the evaluation of cases not complying.

In the Abstract Protocol the prediction is confirmed (see Table 4).

Table 4: Mean scores obtained by participants in the two conditions of the Abstract Protocol.

Age groups	*Complying*	*Not complying*
7-7;9	2.75	2.53
14-14;9	3.07	2.93
21-24;9	3.24	3.07
Overall	3.02	2.84

All groups of participants performed better in the complying condition than in the not complying condition: the difference is statistically significant in each age group (paired T Test; t value ranging from 1.772 to 3.553, p value ranging from <.002 to <.05).

Also, within the Abstract Protocol the ability to evaluate instances not complying with a connective improve with age, as we predicted (ANOVA one-way; F= 7.249, p<.001).

The same results hold in the Pragmatic Protocol (see Table 5).

Table 5: Mean scores obtained by participants in the two conditions of the Pragmatic Protocol.

Age groups	*Complying*	*Not complying*
7-7;9	2.79	2.40
14-14;9	3.21	2.81
21-24;9	3.08	2.92
Overall	3.03	2.71

All groups of participants performed better in the complying condition than in the not complying condition, and the difference is statistically significant in each age group (paired T Test; t value ranging from 1.746 to 3.972, p value ranging from <.001 to <.05). Also within the Pragmatic Protocol the ability to evaluate instances not complying with a connective improve with age, as we predicted (ANOVA one-way; F= 5.994, p<.004).

Thus, MMT predictions hold both within the Abstract and the Pragmatic Protocol.

5. Conclusions

The aim of the experiment was to test the power of MMT in explaining how people represent the meaning of connectives in their mind. The results of the experiment confirm our predictions.

First, the difficulty in comprehending the meaning of a connective depends on the number of mental models it requires. Our results show the following trend of difficulty among connectives in both the Abstract and the Pragmatic Protocol: conjunction is easier than disjunction and bi-conditional, and the latter are easier than conditional.

Second, MMT predicts that, according to the Principle of truth, evaluating instances of compliance is easier than evaluating instances of non-compliance. Our data confirm such a prediction in both contexts. We have argued, in line with MMT, that the evaluation of instances not consistent with a connective leads subjects to err because they have to imagine first the true instances and then to infer the false ones. The corollary prediction that the ability to evaluate instances of non-compliance does increase with the age is also confirmed.

Our results are consistent with the results obtained by Bucciarelli e Johnson-Laird (2001). They investigate reasoning with conditionals within contexts where subjects have to construct instances complying with an assertion (i.e. instances of truth and obedience) and instances not complying with an assertion (i.e. lie and disobedience). Their results show that, while in a selection task the not complying context improves the performance, in a comprehension task participants are better at constructing cases of compliance than cases of non-compliance.

Our results corroborate MMT's predictions in three different age groups (children, adolescents and adults). These results strengthen the theory, which is powerful enough to predict and explain the development of connectives' comprehension. In particular, MMT explains the different difficulty of connectives, the difference in difficulty of comprehension of the same connective (easier in complying conditions, and more difficult in not complying conditions) and, finally, how people represent the meaning of the connectives both within an abstract context and a pragmatic context.

Acknowledgements

This research has been supported by M.U.R.S.T. of Italy, cofinanziamento 1999 (9911314534, Reasoning Processes and Mental Models).

References

Amidon, A. (1976). Children's understanding of sentences with contingent relations: Why are temporal and conditional connectives so difficult? *Journal of Experimental Child Psychology*, *22*, 423-437.

Bara, B.G., Bucciarelli, M., & Lombardo, V. (2001). Model theory of deduction: A unified computational approach. *Cognitive Science*, in press.

Bara, B.G., Bucciarelli, M., Johnson-Laird, P.N., & Lombardo, V. (1994). Mental models in propositional reasoning. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 15-20). Hillsdale, NJ: Lawrence Erlbaum Associates.

Barres, P.E., & Johnson-Laird, P.N. (1997). Why is it hard to imagine what is false? *Proceeding of the Nineteenth Annual Conference of the Cognitive Science Society* (p. 859). Mahwah, NJ: Lawrence Erlbaum Associates.

Bloom, L., Lahey, M., Hood, L., Lifter, K., Fiess, K. (1980). Complex sentences: Acquisition of syntactic connectives and the semantic relations they encode. *Journal of Child Language*, *7*, 235-261.

Braine, M.D.S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review*, *85*, 1-21.

Braine, M.D.S. (1990). The «natural logic» approach to reasoning. In W.F. Overton (Ed.), *Reasoning, necessity, and logic*. Hove, UK: Lawrence Erlbaum Associates.

Braine, M.D.S. (1998). Steps towards a mental predicate logic. In M.D.S. Braine & D.P. O'Brien (Eds.), *Mental logic*. Mahwah, NJ: Lawrence Erlbaum Associates.

Braine, M.D.S., & Rumain, B. (1983). Logical reasoning. In J.H. Flavell & E.M. Markman (Eds.), *Carmichael's handbook of child psychology, Vol. 3. Cognitive Development*. Fourth Ed. New York: Wiley.

Bucciarelli, M., & Johnson-Laird, P.N. (2001). Is there an innate module for deontic reasoning? In J. Garcia-Madruga, N. Carriedo & M.J. Gonzalez-Labra (Eds.), *Mental models in reasoning*. Madrid: UNED.

Cheng, P.W., & Holyoak, K.J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, *17*, 391-416.

Cheng, P.W., Holyoak, K.J., Nisbett, R.E., & Oliver, L.M. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology*, *18*, 293-328.

Cowan, N. (1997). The development of working memory. In N. Cowan (Ed.), *The development of memory in childhood. Studies in developmental psychology*. Hove, UK: Psychology Press.

Griggs, R.A., & Cox, J.R. (1982). The elusive thematic materials effect in the Wason selection task. *British Journal of Psychology*, *73*, 407-420.

Inhelder, B. & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. London: Routledge, Chapman & Hall.

Johansson, B. S., & Sjolin, B. (1975). Preschool children's understanding of the coordinators "and" and "or". *Journal of Experimental Child Psychology*, *19*, 233-240.

Johnson-Laird, P.N. & Barres, P. (1994). When 'or' means 'and': A study in mental models. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 475-478). Hillsdale, NJ: Lawrence Erlbaum Associates.

Johnson-Laird, P.N. & Byrne, R.M.J. (1991). *Deduction*. Hove, UK: Lawrence Erlbaum Associates.

Johnson-Laird, P.N. (1983). *Mental models: Towards a cognitive science of language, and consciousness*. Cambridge, UK: Cambridge University Press.

Johnson-Laird, P.N., Byrne, R.M.J., & Schaeken, W. (1992). Propositional reasoning by model. *Psychological Review*, *99*, 418-439.

Piaget, J. (1953). *Logic and psychology*. Manchester, UK: Manchester University Press.

Pollard, P. (1981). The effect of thematic content on the Wason selection task. *Current Psychological Research*, *1*, 21-30.

Rips, L.J. (1990). Reasoning. *Annual Review of Psychology*, *41*, 85-116.

Staudenmayer, H., & Bourne, J.R.L.E. (1977). Learning to interpret conditional sentences: A developmental study. *Developmental Psychology*, *13*, 6, 616-623.

Taplin, J.E. (1971). Reasoning with conditional sentences. *Journal of Verbal Learning and Verbal Behavior*, *10*, 219-225.

Towse, J. N., Hitch, G. J. & Hutton, U. (1998). A reevaluation of working memory capacity in children. *Journal of Memory and Language*, *39*, 195-217.

A Selective Attention Based Method for Visual Pattern Recognition

Albert Ali Salah (SALAH@Boun.Edu.Tr)
Ethem Alpaydın (ALPAYDIN@Boun.Edu.Tr)
Lale Akarun (AKARUN@Boun.Edu.Tr)
Department of Computer Engineering; Boğaziçi University,
80815 Bebek Istanbul, Turkey

Abstract

Parallel pattern recognition requires great computational resources. It is desirable from an engineering point of view to achieve good performance with limited resources. For this purpose, we develop a serial model for visual pattern recognition based on the primate selective attention mechanism. The idea in selective attention is that not all parts of an image give us information. If we can attend to only the relevant parts, we can recognize the image more quickly and using less resources. We simulate the primitive, bottom-up attentive level of the human visual system with a saliency scheme, and the more complex, top-down, temporally sequential associative level with observable Markov models. In between, there is an artificial neural network that analyses image parts and generates posterior probabilities as observations to the Markov model. We test our model on a well-studied handwritten numeral recognition problem, and show how various performance related factors can be manipulated. Our results indicate the promise of this approach in complicated vision applications.

Introduction

Primates solve the problem of visual object recognition and scene analysis in a serial fashion with *scanpaths* (Noton & Stark, 1971), which is slower but less costly than parallel recognition (Tsotsos, Culhane, Wai, Lai, Davis, Nuflo, 1995). The idea in selective attention is that not all parts of an image give us information and analysing only the relevant parts of the image in detail is sufficient for recognition and classification.

The biological structure of the eye is such that a high-resolution fovea and its low-resolution periphery provide data for recognition purposes. The fovea is not static, but is moved around the visual field in saccades. These sharp, directed movements of the fovea are not random. The periphery provides low-resolution information, which is processed to reveal salient points as targets for the fovea (Koch & Ullman, 1985), and those are inspected with the fovea. The eye movements are a part of overt attention, as opposed to covert attention which is the process of moving an attentional *'spotlight'* around the perceived image without moving the eye.

In the primate brain, information from the retina is routed through the lateral geniculate nucleus (LGN) to the visual area V1 in the occipital lobe. The *'what'* pathway, also known as the *ventral* pathway for anatomical reasons, goes through V4 and inferotemporal cortex (IT). The *'where'* pathway, or the *dorsal* pathway, goes into the posterior parietal areas (PP) (Ungerleider & Mishkin, 1982). The ventral pathway is crucial for recognition and identification of objects, whereas the dorsal pathway mediates the location of those objects. We should note that although recent findings point towards a distinction between perception and guidance of action (Crick & Koch, 1990) instead of a distinction between different types of perception, the issue is not resolved in favour of a specific theory (Milner & Goodale, 1995).

The serial recognition process gathers two types of information from the image: The contents of the fovea window, and the location to which the fovea is directed. We call these *'what'* and *'where'* information, respectively (Ungerleider & Mishkin, 1982). The object is thus represented as a temporal sequence, where at each time step, the content of the fovea window and the fovea position are observed.

Recurrent multi-layer perceptrons were used to simultaneously learn both the fovea features and the class sequences (Alpaydın, 1996). Other techniques are explored in the literature to apply the idea of selective attention to classification and analysis tasks (Itti, Koch, Niebur, 1998; Rimey & Brown, 1990). Our approach is to combine a feature integration scheme (Treisman & Gelade, 1980) with a Markov model (Rimey & Brown, 1990).

We use handwritten numeral recognition to test our scheme. In our database (UCI Machine Learning Repository, Optdigits Database), there are ten classes (numerals from zero to nine) with 1934 training, 946 writer-dependent cross-validation, 943 writer-dependent and 1797 writer-independent test cases. Each sample is a 32 32 binary image which is normalized to fit the bounding box. There are parallel architectures to solve this problem in the literature (Le Cun, Boser, Denker, Henderson, Howard, Hubbard, Jackel, 1989), and they have good performance, but our aim is to design a scalable system which is applicable to problems where the input data is high-dimensional (e.g. face recognition), or not of fixed size (e.g. recognizing words in cursive handwriting). Implementing a parallel scheme with good performance is not trivial in such cases.

This paper is organized as follows: We first describe our model and its three levels. Then we report our simulation results. In the last section we summarize and indi-

cate future directions.

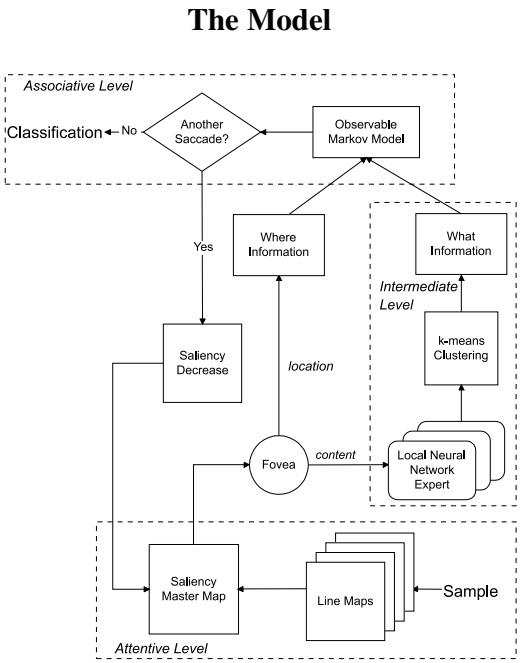


Figure 1: The selective attention model for visual recognition.

The block diagram of the system we propose is given in Fig. 1. It is composed of the *attentive level* that decides on where to look, the *intermediate level* that analyses the content of the fovea, and the *associative level* that integrates the information in time.

Attentive Level

In the first step of the model, the bottom-up part of the visual system is simulated. We work on 12 12 downsampled images to simulate a low-resolution resource. This slightly decreases the classification accuracy, but speeds up the computation considerably. Convolving the digit image with 3 3 line orientation kernels produces four line orientation maps in 0 , 45 , 90 and 135 angles. These are combined in a saliency master map, which indicates the presence of aligned lines on the image.

Line orientations are detected by different primitive mechanisms in the visual cortex, operating in coarse, intermediate and fine scales (Foster & Westland, 1998). We can also talk about simple, complex and hypercomplex cell structures in the visual cortex, that deal with increasing levels of complexity and decreasing levels of resolution. In constructing the saliency map, we use the simplest set of features to decrease the computational cost. Our experiments showed that adding other feature detectors like corner maps, Canny edge detector, and further line orientation maps in higher resolutions increased the classification accuracy only slightly, whereas the increase in the computational cost was significant.

The saliency map indicates the interesting spots on the image. We simulate the fovea by moving a 4 4 window over the 12 12 downsampled image. The saliency values of the visited spots and their periphery are decreased, and these spots are not visited again. This process has a biological counterpart: Once neurons attuned to detect a specific feature fire in the brain, they are temporarily inhibited. Subsequently, subjects respond slower to previously cued locations (Klein, 2000).

Intermediate Level

The simulation of shifts of attention should provide us with *'what'* and *'where'* information, but we want them to be sufficiently quantized to be used in the associative level. We divide the image space into uniform regions, in effect, performing a quantization on the location information. We use a second set of overlappping windows to reduce the effect of window boundaries, as shown in Fig. 2. We obtain a time-ordered sequence of visited regions after the simulation of shifts. This constitutes the *'where'* stream for the particular sample (Fig 3).

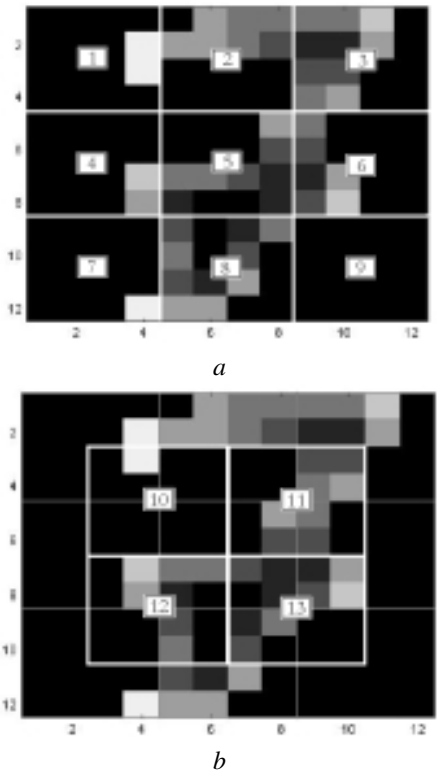


Figure 2: Regions of the downsampled image. (a) The uniform regions. (b) The additional, overlapping regions. Notice how the corner at the intersection of 5^{th}, 6^{th}, 8^{th}, and 9^{th} regions are missed in those regions, but captured clearly in the 13^{th} region.

As fovea contents, we extract 64-dimensional real-valued vectors. These vectors are produced by concate-

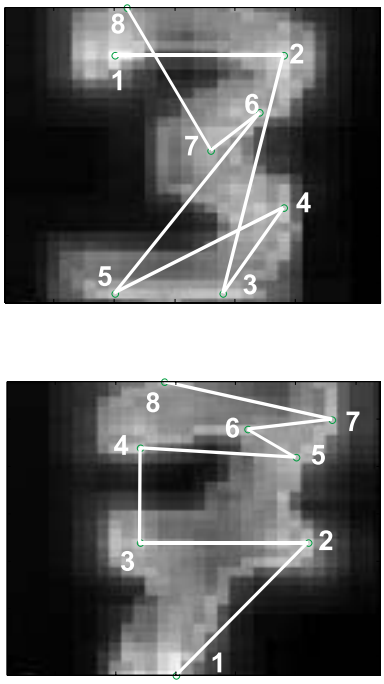


Figure 3: The saliency master maps of two examples from the Optdigits database. Locations with high intensity indicate high saliency values. The locations visited by the fovea are connected with a line, and enumerated in the order they are visited.

nating the corresponding 4 4 windows on the line maps. We prefer using the concatenated line maps to inspecting the original bitmap image, because the line maps indicate the presence of features more precisely. Furthermore, since they were constructed in the attentive level, they come at no additional cost. In any case, we need a vector quantization on the fovea contents before passing them to the associative level.

In order to efficiently quantize this information, we train artificial neural network experts at each region of the image. The experts are single-layer perceptrons (SLP) that are trained in a supervised manner (Bishop, 1995). Their input is the 64-dimensional fovea content vector. The output of the experts are 10-dimensional class posterior probability vectors, which are then clustered with k-means clustering (Duda & Hart, 1973) to obtain the *'what'* information stream. We select single-layer perceptrons over multi-layer perceptrons for a number of reasons. Multi-layer perceptrons overlearn the training data quickly, and perform worse on the cross-validation set. The number of parameters we need to store for the multi-layer perceptron is larger, and the training time is significantly higher. These properties make the single-layer perceptron the better choice of expert in the final model.

Associative Level

In the associative level, the two types of quantized information are combined with a discrete, observable Markov model (OMM) (Rabiner 1989). We treat the regions visited by the fovea as the states of a Markov model, and the quantized output of the local artificial neural network experts as the observations from each state. We simulate eight shifts for each sample in the training set, obtain the *'where'* and *'what'* streams, and adjust the probabilities of the single Markov chain of the corresponding class to maximize the likelihood of the training data. Training an observable Markov model is much faster than training a Hidden Markov Model.

In the observable model, the model parameters are directly observed from the data. Since we know the states, we can count the state transitions, and normalize the count to find the state transition probabilities a_{ij}, as well as the initial state distribution probabilities π_i. Similarly, we count the occurrences of the observation symbols (quantized outputs of the local neural networks) at each state, and normalize them to find the observation symbol probability distribution $b_j\ k$:

$$\pi_i \qquad \frac{\#\text{ of times in } S_i \text{ at time } t \quad 1}{\#\text{ of observation sequences}} \tag{1}$$

$$a_{ij} \qquad \frac{\#\text{ of transitions from } S_i \text{ to } S_j}{\#\text{ of transitions from } S_i} \tag{2}$$

$$b_j\ k \qquad \frac{\#\text{ of times in } S_j \text{ observing } v_k}{\#\text{ of times in } S_j} \tag{3}$$

Finding the probability of the observation sequence is much simpler in the observable Markov model, since the states are visible. We just multiply the corresponding state transition probabilities and the observation probabilities:

$$P\ O\ S\ \lambda \qquad \pi_{S_1} b_{S_1}\ v_1 \prod_{i\ 2}^{n} a_{S_i\ 1 S_i} b_{S_i}\ v_i \tag{4}$$

where S is the state sequence, O is the observation sequence, and $\lambda \qquad \pi_i\ a_{ij}\ b_j\ k$ stands for the parameters of the Markov model. $i\ j \quad 1\ N$ are indices for states, $k \quad 1\ M$ is the index for the observation symbols.

The Markov model is trained with a limited training set, and if the number of states and observation symbols is large, there will be connections that are not visited at all. Since the model is probabilistic, having a transition or observation probability of zero is not desired. Instead, the transitions that have not occurred within the training set should have a low probability in the model. This is what we do in the post-processing stage. We scan the probabilities of the trained Markov model, and replace all probabilities lower than a threshold (0 001) with the threshold value. Then we normalize the probabilities once more. This is a simple and fast procedure that achieves the desired effect.

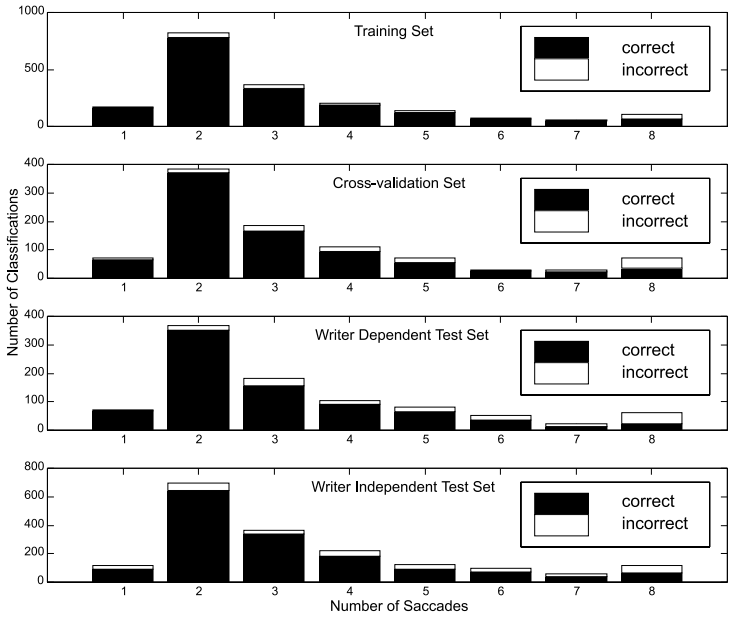


Figure 4: Dynamic fovea simulation results. These are the histograms of the number of correctly and incorrectly classified digits after each shift. See also Figs. 5 and 6.

We have also tried Hidden Markov Models where the states are not visible and where the concatenated *where-what* information is the observation, but this structure performed worse than the observable Markov model.

Dynamic Fovea

One important advantage of using a Markov model is the ease with which we can control the number of shifts necessary for recognition. In the training period, our model simulates eight shifts, which is set as the upper bound for this particular application. After each shift, the Markov model has enough information to give a posterior probability for each class. We may calculate the probability $\alpha_t\ c$ of the partial sequence in the Markov model, which reflects the probability of the sample belonging to a particular class, given the *'where'* and *'what'* information observed so far. Using Eq. 4, we have

$$\alpha_t\ c \quad P\ O_1 \quad O_t\ S_1 \quad S_t\ \lambda_c \tag{5}$$

where $O_1 \quad O_t$ is the observation sequence up to time t, $S_1 \quad S_t$ is the state sequence, and λ_c are the parameters of the Markov model for class c.

We can use this probability to stop our shifts whenever we reach a sufficient level of confidence in our decision. Let us define $\alpha_t\ c$, the posterior probability for class c at time t:

$$\hat{p}\ c\ O_1 \quad O_t\ S_1 \quad S_t \quad \alpha_t\ c \quad \frac{\alpha_t\ c}{\sum_{j\ 1}^{K} \alpha_t\ j} \tag{6}$$

Let τ be the threshold we use as our stopping criterion:

$$\alpha_t\ c \quad \tau \tag{7}$$

where the value of τ is in the range [0,1]. If we assume that absolute certainty is not reached anywhere in the model and $\alpha_t\ i$ is always below 1 0, selecting $\tau \quad 1\ 0$ is equivalent to treating all samples as equally difficult and doing eight shifts. Conversely, selecting $\tau \quad 0$ is equivalent to looking at the first salient spot and classifying the sample.

Selecting a large value for τ trades off speed for accuracy. With a well selected value, we devote more time for difficult samples, but recognize a trivial sample in a few shifts.

Results

In this section we present our simulation results. We give additional information about the techniques we employ in subsections.

Local Experts

Implementing local artificial neural network experts both increases the classification accuracy and decreases the complexity and classification time. The single-layer perceptron returns a 10-dimensional vector from a 64-dimensional linemap image. Since it is trained in a supervised manner, it provides more useful information for classification to the later Markov model, as our experiments indicated.

Dynamic Fovea Simulation

When we simulate the dynamic fovea with a fixed threshold of τ 0 95, we get 85 67 per cent classification accuracy with 5 46 per cent standard deviation on the writer-dependent test set. The average number of shifts is 3 33, which corresponds roughly to seeing one thirds of the image in detail. This justifies our claim that analyzing only a small part of the image is enough to recognize it. On the writer independent test set, the classification accuracy is 84 63 per cent, with a standard deviation of 7 58 per cent, and the average number of shifts is 3 37 (See Fig. 4 for the histograms depicting the distribution of classifications over the shifts). We are doing less than half the number of shifts we were doing, but the performance decrease is less than a standard deviation.

The advantages of simulating a dynamic fovea become apparent when we inspect Figs. 5 and 6. The accuracy of classification increases when we increase the threshold, because a higher threshold means making more shifts to get a more confident answer. A lower threshold means that a quick response is accepted. What happens is that the average number of shifts increase sharply if the threshold is set to a value very close to 1.0. In this case, the classifier cannot exceed the threshold probability with eight shifts, and selects the highest probability class, without doing any more shifts. This is the reason behind the relatively high number of correct and incorrect classifications after eight shifts in Fig. 4.

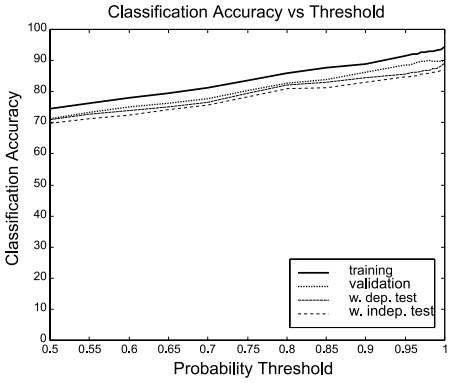


Figure 5: Accuracy vs threshold value in dynamic fovea simulation

Simulation Results

We summarize the results we obtain in Table 1. The first column of the table shows the method employed. The successive columns indicate the classification accuracy and its standard deviation on the training, cross-validation, writer-dependent test and writer-independent test sets.

In the first two rows, we do eight shifts, generate the posterior probabilities of classes by the local artificial neural network experts and take a vote without treating them as a sequence. Soft Voting takes into account the 10-dimensional outputs of the experts instead of a single class code. Comparing the results with the OMM results show that the order information which is lost during voting but used in OMM is useful. Another observation is that the post-processing method we use increases the performance by one standard deviation, which is a significant increase.

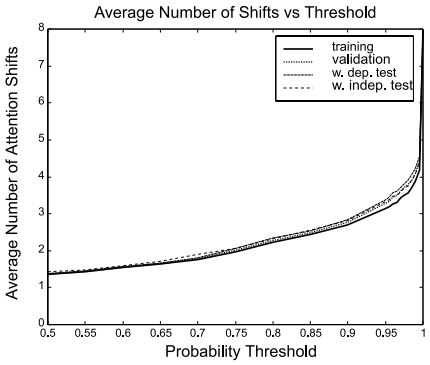


Figure 6: Average number of shifts vs threshold value in dynamic fovea simulation

The dynamic fovea simulation has a lower classification accuracy, but it only needs 3 2 shifts on the average, instead of the previous eight.

Finally, the last row indicates the accuracy of an all-parallel scheme. We use a multi-layer perceptron (MLP) with 32 32 binary input and 10 hidden units. Although the MLP has a good accuracy in this problem, it is not scalable due to the curse of dimensionality.

Conclusions and Future Work

The selective attention mechanism exploits the fact that real images often contain vast areas of data that are insignificant from the perspective of recognition. A low-resolution, downsampled image is scanned in parallel to find interesting locations through a saliency map, and complex features are detected at those locations by means of a high-resolution fovea. Recognition is done serially as the location and feature information is combined in time. By keeping the parallel part of the method simple, we can speed-up the recognition process considerably.

Our tests have demonstrated that an observable Markov model may replace an HMM for the two-pathway selective attention model. The observable scheme is easier to train and use, and performs better. The dynamic fovea simulation reveals further benefits of serializing the recognition process. We can control the time we spend on an image, and differentiate between simple and confusing images. This is a desirable property in a classifier, since it allows us to apply more reliable and costly methods to the confusing samples if we wish. It also reduces the average recognition time, but it

Table 1: Summary of Results

Method	Performance			
	Training	**Validation**	**Writer Dep. Test**	**Writer Indep. Test**
SLP+Simple Voting	86.74(9.90)	85.92(9.39)	64.51(25.62)	62.66(25.74)
SLP+Soft Voting	93.85(4.47)	91.25(7.07)	74.35(27.66)	73.89(26.67)
OMM+SLP	95.32(3.72)	83.98(15.37)	84.42(14.94)	80.92(16.24)
OMM+SLP + post-processing	94.37(3.33)	90.07(7.92)	89.73(8.68)	87.37(8.73)
Dynamic fovea	91.41(4.56)	88.47(7.98)	85.67(5.46)	84.63(7.58)
MLP	99.92(0.12)	97.45(0.28)	97.25(0.42)	94.54(0.21)

must be remembered that the construction of the saliency map is necessary for all samples. Although we reduce the time complexity of the associative level by half, the overall gain is less than that.

Our attempt to classify digits may be seen as a toy problem, since the ratio of the fovea area to the image is not high enough to demonstrate the benefits of our model. Although the accuracy is lower than the state-of-the-art parallel approaches in the literature (e.g. the MLP result in Table 1), the selective attention mechanism is much more appropriate for applications where parallel processing is too cumbersome to use, and the number of input dimensions is high.

We are planning to employ our model in a more difficult task, such as face recognition, where an all-parallel classifier, like the MLP, would be unnecessarily complex; in a face, small regions of the face like eyes, nose, mouth give us information. The saliency scheme has to be modified for this purpose, as facial features necessitate different and more complex feature detectors. The fovea size also needs to be adjusted for the specific task.

Acknowledgments

This work is supported by Boğaziçi University Research Funds 00A101D.

References

Alpaydın, E. (1996). Selective Attention for Handwritten Digit Recognition. In D.S. Touretzky, M.C. Mozer, & M.E. Hasselmo (Eds.), *Advances in Neural Information Processing Systems 8*, 771-777.

Bishop, C.M. (1995). *Neural Networks for Pattern Recognition.* Oxford University Press.

Crick, F., & Koch, C. (1990). Towards A Neurobiological Theory Of Consciousness. *Seminars in the Neurosciences*, 2, 263-275.

Duda, R., & Hart, P. (1973). *Pattern Classification and Scene Analysis.* New York: John Wiley & Sons.

Foster, D.H., & Westland, S. (1998). Multiple Groups of Orientation-selective Visual Mechanisms Underlying Rapid Oriented-line Detection. *Proc. Royal Society London*, 265, 1605-1613.

Itti, L., Koch, C., & Niebur, E. (1998). A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20, 11.

Klein, R.M. (2000). Inhibition of Return. *Trends in Cognitive Science*, 4(4), 138-147.

Koch C., & , Ullman, S. (1985). Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. *Human Neurobiology*, 4, 219-227.

Le Cun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., & Jackel, L.D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1, 4, 541-551.

Milner, A. D., & Goodale, M. A. (1995) *The Visual Brain in Action.* Oxford University Press.

Noton, D, & Stark, L. (1971). Eye Movements and Visual Perception. *Scientific American*, 224, 34-43.

Rabiner, L.R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. IEEE*, 17, 2.

Rimey, R.D., & Brown, C.M. (1990). Selective Attention as Sequential Behavior: Modeling Eye Movements with an Augmented Hidden Markov Model. (Tech. Rep. TR-327). Computer Science, University of Rochester.

Treisman, A.M., & Gelade, G. (1980). A Feature Integration Theory of Attention. *Cognitive Pscychology*, 12, 1, 97-136.

Tsotsos, J.K., Culhane, S.M., Wai, W.Y.K., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling Visual Attention via Selective Tuning. *Artificial Intelligence*, 78, 507-545.

UCI Machine Learning Repository, Optdigits Database, prepared by E. Alpaydın and C. Kaynak. ftp://ftp.ics.uci.edu/pub/machine-learning-databases/optdigits.

Ungerleider, L.G., & Mishkin, M. (1982). Two cortical visual systems. In D.J. Ingle, M.A. Goodale, & R.J.W. Mansfield (Eds.), *Analysis of visual behavior*, MIT Press.

Solving arithmetic operations: a semantic approach

Emmanuel Sander (sander@univ-paris8.fr)
University of Paris 8, Department of Psychology, 2 Rue de la Liberté
Saint-Denis, 93526 France

Abstract

Systematic errors observed when solving arithmetic operations are often considered as being procedural. Rules induced by the learner and the errors committed are viewed as resulting by use of general problem solving methods. In this study, in the case of solving column subtractions, we show that some errors may be semantic: they are due to analogies involving different sources that guide the interpretation of both operations and procedures already learned. Two sources have been identified: (i) subtracting considered as removing something and (ii) subtracting considered as covering a distance between two elements. Results of the 2 experiments reported here show that (i) not only all of the predicted semantic errors were observed among beginner children, but also that (ii) semantic errors were still observed among more advanced learners, in a decreasing proportion for higher level of instruction. These results support the idea that (i) semantic aspects have a major influence in the learning process, and (ii) that this kind of errors still intervene in the learning process even if the procedural aspects become more influent (advanced students). These results suggest that a procedural approach might be articulated with a semantic approach.

Introduction

Many studies show that some errors made when solving arithmetic operations (e.g. Brown & Burton, 1978; Sleeman, 1982), and notably when solving column subtractions (e.g. Young & O'Shea, 1981, VanLehn, 1982, 1983, 1987, 1990), are systematic in nature. As it has been observed, some errors are quite stable, both for within or between subjects' measures. This led researchers to reject the idea that all errors are calculation errors or due to lack of attention, and therefore, to look for mechanisms that lead to these errors.

As it has been noticed by Ohlson and Rees (1991), most of the investigators in this field focused attention on procedural mechanisms. The most prominent view for column subtractions has been developed by VanLehn and his colleagues (e.g. Brown & VanLehn, 1980; VanLehn, 1982, 1990) and has promoted the repair theory that has been implemented in the SIERRA Model (VanLehn, 1987, 1990).

In this view, conjunction of two mechanisms may lead to error production. First, learning consists on inducing rules in a syntactic way from lessons composed of solved examples, using general problem solving methods. Second, errors result from application of problem solving heuristics to overcome the impasses encountered when solving a new operation.

This theory can be qualified as procedural because the only parameters which intervene in the model are (i) knowledge about specific arithmetic facts (for instance the fact that 2 is smaller than 5 or 7 – 4 = 3), (ii) heuristics for deriving rules from previously solved examples, (iii) heuristics for solving an impasse situation which has been encountered when previously learned rules have been applied to a new situation. One characteristic of a procedural approach is that the children's interpretation of both the whole operation and the procedure is not taken into account.

Framework

Although VanLehn (e.g. 1990) shows convincing evidence of procedural errors, the exhaustiveness of his description could be questioned. In this paper, we intend to provide evidence that a procedural approach doesn't provide an explanation for some of the systematic errors observed and that articulating this approach with a semantic point of view may increase the range of explained errors. This will show as well that in a procedural situation par excellence, such as solving column subtractions that could be solved in a purely syntactical manner (Resnick, 1982), semantic aspects still influence the solving process. This point has some implications on problem solving mechanisms.

Some work focused on the influence of semantic aspects in solving column subtractions (Carpenter, Franke, Jacobs & Fennema, 1996; Fuson, 1986; Fuson & Briars, 1990; Hiebert & Wearne, 1996; Resnick, 1982; Resnick & Omanson, 1987). Through the use of analogies and/or concrete materials, these studies evaluated the influence of a teaching method and aimed on helping children to understand some of the conceptual background of column subtractions solving methods. The efficiency of these methods is rather variable. In this work, we do not adopt any position in the debate concerning virtues of providing conceptual background versus teaching procedures as if they were arbitrary, but we shall show that semantic aspects are involved spontaneously even if they are not an explicit part of the teaching method. They are expressions of a basic analogical transfer mechanism that attributes a

meaning to a given situation. Such semantic influences have been identified for arithmetic operations by Fischbein (Fischbein, Deri, Nello & Marino, 1985 ; Fischbein, 1989) with the notion of tacit models. These are simple structural entities of a concrete and practical nature, which control the course of the reasoning process and are specific cases of analogy sources (Fischbein, 1987; Sander, 2000). Fischbein and his colleagues worked mainly with word problems and didn't extend their view to procedural situations such as column subtractions.

In the case of solving column subtraction, we hypothesize that errors are not necessarily due to a repair used in an impasse situation. Rather, they are sometimes a direct consequence of the interpretation of a learned procedure: the child applies the procedures according to her/his interpretation and the wrong result might be predicted from this irrelevant interpretation. In order to deal with the target situation (the column subtraction), s/he refers to a source knowledge, which is the knowledge associated with this new subtracting situation. In fact, we hypothesize that the semantic errors observed in column subtractions result from an analogical transfer in situations for which the source is non adequate. Thus, the errors are consequences of a negative transfer. We consider two sources of analogy for solving column subtractions: 'remove' and 'distance'. Both of them can be evoked spontaneously by the children or be due to the teaching method.

In the 'remove' view, subtracting is seen as taking a part out of a whole: the whole and the part are the two quantities and the result is what is left. In the 'distance' view, subtracting is seen as going from a given point to another: the departure and the arrival points are the two values and the result is the distance between them. These interpretations are valid for each value of the whole operation, but the negative transfer is due to its extension to each column and to each digit of the operation. If the operation is interpreted through these sources, the resulting errors can be predicted by these hypothesized sources. Using the terminology of VanLehn (1990), those errors are described in Table 1.

It can be noticed that most of the errors might result from distance or remove interpretations and thus, verbal reports might be useful for identifying the source. It can also be noticed that two errors (Diff 0-N=N and Diff 0-N=0) might sometimes be particular cases of other errors (respectively Smaller from Larger and Zero instead of Borrow) but they might also be specific to the cases involving zero, which could be identified either by verbal reports or by the presence of one error when the other is absent.

From our point of view, when starting to learn, children build interpretations of the operation and of the procedures through analogical transfer mechanism. This

Table 1: Definition and interpretation of semantic errors

Smaller from Larger (e.g. 457-168=311)
Definition: The smaller digit from each column is subtracted from the larger one wherever it is situated.
Remove interpretation: When a part is removed from a whole, the part is always smaller than the whole.
Distance interpretation: As a distance is symmetrical, the distance from the smaller to the larger is equal to the distance from the larger to the smaller.
Zero instead of Borrow (e.g. 457-168=300)
Definition: A zero is written instead of borrowing.
Remove interpretation: If what has to be removed is more than what is available, then it is removed and a zero is left. The impossibility of removing a quantity larger than the whole might also be marked by a zero.
Distance interpretation: In this case, distance is considered as oriented. If one considers that it is only possible to go upward, the distance is zero when the departure point is situated after the arrival point.
Blank instead of Borrow (e.g. 457-168=3 _).
Definition: Nothing is written instead of borrowing.
Remove interpretation: No answer is given in the corresponding column to signal the impossibility of removing something larger than what is actually there.
Distance interpretation: The impossibility to go backward is marked by a non-answer to the corresponding column.
Stutter subtract (e.g. 457-3=124)
Definition: The last digit of the same line takes the place of a blank.
Remove interpretation: Two quantities are needed when removing a part, thus the missing quantity is replaced by the closer one.
Distance interpretation: Departure point and arrival point are both needed to go from one place to another. Thus, the missing point is replaced by the closer one.
Diff 0-N=N (e.g. 400-168=368)
Definition: If one of the upward digits is zero, the downward digit is written as the result.
Remove interpretation: Nothing can be taken from zero so the original value is unchanged.
Distance interpretation: None.
Diff 0-N=0 (e.g. 400-168=300)
Definition: If one of the upward digits is a zero, zero is written as the result.
Remove interpretation: The impossibility of taking something from zero is marked by a zero result.
Distance interpretation: None.
Diff N-0=0 (e.g. 457-100=300)
Definition: If one of the downward digits is zero, the zero is written as the result.
Remove interpretation: The impossibility of taking zero from a quantity is marked by a zero result.
Distance interpretation: None

extends the range of application of the learned procedures.

For instance, with the "remove" and "smaller from larger" interpretations of a procedure, the child will extend to "3 – 6" what has been learned from "6 – 3". S/he will consider that they are both cases of "removing a part from a whole", and that the place of the digit (either upward or downward) is not relevant because the part is necessarily the smaller number and the whole is necessarily the larger number. With acquisition of new procedures, semantic influence will persist but will decrease since specific learned procedures will narrow the extension of semantic interpretation. At the same time, procedural errors, as it has been established (e.g. VanLehn, 1990) will be developed.

Thus, we make the following hypotheses:

(i) At the beginning of learning, children will mostly make semantic errors resulting from analogical transfer with the hypothesized sources.

(ii) Semantic errors will persist even among more experienced children

(iii) The relative proportion of semantic errors of the total number of errors decreases as child level increases.

The aim of the first experiment is to test the first hypothesis and the aim of the second experiment is to test hypotheses (ii) and (iii).

Experiment 1

Subjects

Subjects were 50 grade 2 children who begun studying how to subtract but haven't began studying how to borrow. In accordance with the ministerial directives, the teaching methods in the classrooms were focusing on the procedural aspects and not on the conceptual backgrounds.

Material and procedure

Children had to solve collectively in the classroom, and without any time limit, 20 subtractions used by VanLehn (1982), which allowed identification of a large variety of errors. The instruction was: "You have to solve 20 subtractions. Do the best you can. Take all the time you need." Since 17 of the 20 required borrowing, a high rate of wrong results was expected.

This situation is quite original in the didactical field, since students are usually tested on contents that they have supposedly already studied. However, it is standard in problem solving paradigms that specific knowledge about the problems is often considered as a bias that has to be avoided. In fact, to support the existence of a general cognitive mechanism involved in this situation, we used an usual problem solving paradigm in a school situation.

Furthermore, 14 of the 50 children were randomly chosen and were tested again the next week after the first test. They were asked to solve the same 20 operations for a new test and to explain the way by which they solved them. Their verbal reports were recorded.

Results

Protocols of 8 children were excluded from the data because of their use of the borrowing procedure, probably learned at home.

Quantitative results

Despite the fact that the subjects learned only how to solve 15% of the operations (3 out of 20), they answered 84.5% of the operations in average. Only one subject answered the 3 operations corresponding to what he had actually learned.

Table 2 displays the results. Dominant errors were distinguished from partial errors, depending on their rate of occurrence within a same protocol.

Table 2: Rate of semantic errors

Error	Dominant	Partial	Total
Smaller from Larger	47.6%	19.0%	66,6%
Zero instead of Borrow	38.1%	16.7%	54.8%
Blank instead of Borrow	4.8%	2.4%	7.2%
Stutter Subtract	9.5%	11.9%	21.4%
Diff 0-N=N	50.0%	7.1%	57.1%
Diff 0-N=0	38.1%	16.7%	54.8%
Diff N-0=0	9.5%	0.0%	9.5%

As it can be noticed, some errors are very usual: 4 types of errors are observed for more than half of the subjects. Furthermore, all the hypothesized errors were observed. Few non hypothesized errors were observed, but only for a minority of the subjects (9.5%): Small-Large = Small; Small-Large = Large; N-N=N.

These results have to be contrasted with predictions of procedural approaches. As a matter of fact, VanLehn (1990), observed all those errors but did not generate a large part of them with SIERRA, that generated only "Smaller from Larger" and "Blank instead of Borrow" from this list. In other words, errors as frequent as "Zero instead of Borrow", "Diff 0-N=N" and "Diff 0-N=0" were not produced by SIERRA. All in all, 76.2% of the children revealed errors that were not predicted by this model, when only 9.5% of them revealed errors not predicted by the semantic approach.

Simulations

3 levels of simulation were performed. At the first level, each protocol was associated with a list of non competitive errors that were observed for this protocol: for instance, Diff 0-N=0 and Diff 0-N=N could not be associated with the same protocol because they lead to different results. The actual results were compared with the ones obtained when applying the identified errors to

the operation. No calculation error was allowed for explaining differences.

At the second level, calculation errors were taken into account: a difference of plus or minus 1 was accepted if this was not corresponding to another systematic error.

At the third level, all the errors observed in the protocol were taken into account even if some of them were competitive.

Thus, several results could be compatible with the same simulation and calculation errors were taken into account as well. The results are presented in Table 3.

Table 3: Simulation with semantic errors

	Sim 1	Sim 2	Sim 3
Rate of prediction	77.9%	81.3%	89.8%

Verbal reports

Verbal reports were analyzed in the following way.

First, various expressions were identified as cues for specifying a source. For instance, "I have 3 and I remove 4" was indicating a 'remove' source since 'I have' was considered as referring to a quantity and 'remove' referring to the conception of taking part of a whole. "Going from 0 to 7…" indicates a distance source because of the reference of going from one place to another. This analysis of verbal reports showed that each child was using these kinds of expressions, which support the idea that the operation is actually interpreted in terms of taking part of a whole or going from one place to another.

Second, explanations for each error were compared to the predicted interpretations. The explanations produced were consistent with the expected ones.

Leaving apart the debate concerning verbal protocols, we would like to point out that verbal reports here coincide with the expected interpretations.

Hereafter, few examples are presented.

Subject J (83-44=40; Zero instead of Borrow; remove): "There are 3. We have to remove more than what is there. It remains 0"

Subject A (1564-887=1000; Zero instead of Borrow; remove): "… I have 5 candies in my hand, I cannot take 8 out of them to eat, so I take the 5 and nothing is left…"

Subject H (6591-2697=4106; Smaller from Larger; remove): "I have 1, no I have 7, I remove 1, 6 is left; I have 9, I remove 9, 0 is left; I have 6, I remove 5, 1 is left …"

Subject D (8305-3=5002; Stutter Subtract and Diff 0-N=0; remove): "We put 5 on the fingers, we remove 3 and we notice that 2 are left. 0 can't be removed from any number so it makes 0. And 3, I remove 3, 0 is left. For 8, I remove 3 and 5 are left from the 8."

Subject F (6591-2697=4106; Smaller from Larger; distance): " We count in our head what is missing to go from a given number to the one we need. From 1 to 7, 6 is missing, thus 7 minus 1 is 6; from 9 to 9, 0 is missing; from 0 to 9, 9 is missing; from 5 to 6, 1 is missing …"

Subject B (562-3=231; Stutter Subtract & Smaller from Larger; distance). "3 minus 2 is 1 and then since there is nothing I have to use the 3. From 3 to 6, 3 is needed. From 3 to 5, 2 is needed".

Discussion

As a summary, results of this first experiment support the hypothesis that interpretative aspects are involved in solving column subtractions.

First, despite the fact that children didn't know how to solve most of the problems, they tried to answer nearly all of them. From our point of view, this result supports the idea that what the children learned was interpreted in a conceptual framework that provided solutions for new situations.

Second, all the errors predicted by the semantic perspective were actually observed, and unpredicted errors appeared only seldom, supporting the idea that they resulted from the predicted interpretations. Third, depending on the kind of simulation, the semantic perspective allowed prediction of 77.4% to 89.1% of the errors, supporting the idea that semantic errors are prominent in the beginning of learning. Fourth, the analysis of the verbal reports showed that all the children were referring to the hypothesized sources with the predicted interpretations.

Experiment 2

Subjects

409 children who had already studied subtractions with borrowing participated in this experiment. 158 were children of grade 2 and 251 were grade 3. The teaching method was the same as in experiment 1. Given that errors rate might decrease with learning, a greater number of participants was necessary in this experiment to identify the errors.

Material and procedure

The material and procedure were the same as in experiment 1. 16 participants, randomly chosen among the ones who revealed semantic errors, were selected here for verbal reports.

Results

Quantitative results

As in experiment 1, dominant errors were distinguished from partial errors. 39.4% of the children had at least one semantic error. For 13.0% of them, at least one semantic error was dominant and for the remaining 26.4%, the semantic errors were partial. The results per error are presented in Table 4. Results show that the

semantic errors don't disappear with learning. They are still present, and with a great variety, for more than one third of the children, and they stay dominant for a minority of the participants.

Differences between grade 2 and grade 3 children are significant. 52.5% of the grade 2 versus 31.0% of grade 3 children had at least one semantic error ($Chi2(1)=18.70$; $p<.01$). The difference is also significant for the dominant errors: 25.9% of grade 2 versus 4.8% of grade 3 children had at least one dominant semantic error ($Chi2(1)=38.52$, $p<.01$). These differences do not reflect only the better performance of the grade 3 children but also a decrease in the proportion of semantic errors among the whole range of errors as indicated by the simulations.

Table 4: Rate of error for each semantic error

Error	Dominant	Partial	Total
Smaller from Larger	3.1%	11.3%	14.4%
Zero instead of Borrow	0.5%	5.6%	6.1%
Blank instead of Borrow	0.0%	0.0%	0.0%
Stutter Subtract	0.0%	7.1%	7.1%
Diff 0-N=N	8.7%	14.5%	23.2%
Diff 0-N=0	3.6%	16.7%	12.2%
Diff N-0=0	2.0%	6.6%	8.6%

Simulations

The method is similar to the one used in experiment 1. Table 5 displays the results of the simulations, i.e. the percentage of the errors that the semantic approach can explain. Even if there is a strong discrepancy with the beginners' simulation, the semantic perspective allows to predict within 1/4 to 1/5 of the errors for grade 2 children, depending on the kind of simulation, and about 1/7 of the third grade's errors. The differences between each level (grade 2 no borrow, grade 2 borrow, grade 3) are significant for each simulation (Fisher-Pitman Homogeneity Tests with z values from 2.74, $p=.006$ to 12.07, $p<.0001$).

Table 5: Simulation depending on participant level

	Sim 1	Sim 2	Sim 3
Grade 2 no borrow	77.9%	81.3%	89.8%
Grade 2 borrow	21.7%	22.8%	25.3%
Grade 3	14.1%	14.7%	16.3%

These results support our hypothesis that semantic errors persist even after procedural learning and that their importance decreases among the whole range of errors.

Verbal reports

Analysis of the verbal reports leads to the same conclusion as in experiment 1: all of the children referred to the predicted sources, with the predicted interpretations.

Discussion

The results of experiment 2 are consistent with hypotheses (ii) and (iii), supporting the idea that procedures are interpreted within a conceptual framework that definitely has a decreasing influence, yet does not completely disappear even after learning.

This may suggest that not only two kinds of errors exist (procedural and semantic ones) but also that the influence of one decreases when that of the other increases.

The plausibility of this suggestion is reinforced by some of VanLehn's (1990) results. Indeed, VanLehn illustrates (chapter 7) the performance of SIERRA with 33 systematic errors. SIERRA generates 25 of them. Among these 25, only 2 are considered as semantic by our approach. In contrast, among the 8 that SIERRA failed to generate, 5 are predicted by our approach. Thus, it appears that SIERRA's successes are failures of the semantic approach and vice-versa. This supports the idea of different mechanisms.

General Discussion

In this study, we showed through several converging measures that semantic aspects were involved in a procedural task par excellence: solving column subtractions.

Results are consistent with the idea that interpretative aspects should be taken into account in problem solving situations, as it has already been demonstrated in a body of research with some mathematical word problems (e.g. Bassok & Olseth, 1995; Bassok, Wu, & Olseth, 1995), with puzzle problems (Clément & Richard, 1997; Zamani & Richard, 2000), and in learning devices (Sander & Clément, 1997; Sander & Richard, 1997).

Acknowledgments

Thanks to Jean-François Richard, to Mojdeh Zamani and to Michal Eisenstein for their insightful help during the carrying out of this research and the preparation of this manuscript.

References

Bassok, M., & Olseth, K.L. (1995). Object-based representations: Transfer between cases of continuous and discrete models of change. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21,* 1522-1538.

Bassok, M., Wu, L.L., & Olseth, K.L. (1995). Judging a book by its cover: Interpretative effects of content on problem-solving transfer. *Memory and Cognition, 23*, 354-367.

Brown, J.S., & Burton, R.R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2,* 155-192.

Brown, J.S., & VanLehn, K. (1980). Repair Theory : A generative theory of bugs in procedural skills. *Cognitive Science, 4,* 379-426

Carpenter, T.P, Franke, M.L., Jacobs, V.R., & Fennema, E. (1996). Invention and understanding in the development of multidigit addition and subtraction procedures: A longitudinal study. *Annual meeting of the American Research Association*, New York.

Clément, E., & Richard, J-F. (1997). Knowledge of domain effects in problem representation: the case of Tower of Hanoi isomorphs, *Thinking and Reasoning.*

Fischbein, E. (1987). Intuition in science and mathematics: An Educational Approach. Reider: Dordrecht.

Fischbein, E., (1989). Tacit models and mathematical reasoning, *For the Learning of Mathematics, 9*, 9-14

Fischbein, E., Deri, M., Nello, M. S., & Marino, M. S. (1985). The role of implicit models in solving verbal problems in multiplication and division. *Journal for Research in Mathematics Education, 16,* 3-17

Fuson, K. (1986). Role of representation and verbalization in the teaching of multi-digit additions and subtractions. *European Journal of Psychology of Education, 1*, 35-56.

Fuson, K., & Briars, D.J. (1990). Using base-ten blocks learning/teaching approach for first and second grade place value and multidigit additions and subtraction. *Journal for Research in Mathematics Education, 21,* 180-206.

Hiebert, J., & Wearne, D. (1996). Instruction, understanding, and skill in multidigit addition and subtraction. *Cognition and Instruction*, 251-284.

Olhlson, S, & Rees, E. (1991). The function of conceptual understanding in the learning of arithmetic procedures. *Cognition and Instruction*, 103-180.

Resnick, L.B. (1982). Syntax and semantics in learning to subtract. In T. P. Carpenter, J. M. Moser and T. A. Romberg (Eds.), *Addition and substraction: A cognitive perspective* (pp. 136-155). Hillsdale: Erlbaum.

Resnick, L.B., & Omanson, S.F. (1987). Learning to understand arithmetic. In R. Glaser (Ed.), *Advances in instructional psychology, vol 3* (pp. 41-95). Hillsdale, NJ: Erlbaum.

Sander, E. (2000). *L'analogie, du naïf au créatif: analogie et catégorisation.* Paris: L'Harmattan.

Sander, E., & Clément, E. (1997). The interpretative factors of the part-whole dimension in a problem solving situation. *In proceedings of the 5th European Congress of Psychology*. Dublin: Psychological Society of Ireland.

Sander, E., & Richard, J-F. (1997). Analogical transfer as guided by an abstraction process: The case of learning by doing in text editing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 1459-1483.

Sleeman, D.H. (1982). Assessing competence in basic algebra. In D. Sleeman, and J.S. Brown (Eds.), *Intelligent tutoring systems* (pp. 186-199). New York: Academic Press.

VanLehn, K. (1982). Bugs are not enough: Empirical studies of bugs, impasses and repairs in procedural skills. *Journal of Mathematical Behavior, 3,* 3-71.

VanLehn, K. (1983). The representation of procedures in repair theory. In H.P. Ginsburg (Ed.), *The development of mathematical thinking*. Hillsdale, NJ: Erlbaum.

VanLehn, K. (1987). Learning one subprocedure per lesson. *Artificial Intelligence, 31,* 1-40.

VanLehn, K. (1990). *Mind bugs: origins of procedural misconceptions.* Cambridge, Mass.: MIT Press

Young, R.M., & O'Shea, T. (1981). Errors in children's subtraction. *Cognitive Science, 5,* 153-177

Zamani, M., & Richard, J-F. (2000). Object encoding, goal similarity and analogical transfer. *Memory & Cognition, 28,* 873-886.

Do Perceptual Complexity and Object Familiarity Matter for Novel Word Extension?

Catherine M. Sandhofer (csandhof@indiana.edu)
Department of Psychology, 1101 E. 10th Street
Bloomington, IN 47404 USA

Linda B. Smith (smith4@indiana.edu)
Department of Psychology, 1101 E. 10th Street
Bloomington, IN 47404 USA

Abstract

This paper examines the relationship between shape complexity and familiarity in extending novel adjectives. Previous research has suggested that familiarity with an object's basic level label determines the likelihood that a novel adjective will be extended to new instances. The present results do not support that conclusion. Instead the results suggest that given an adjectival syntactic frame children are likely to extend novel words to other objects of the same material when the objects are simple in shape. This result suggests that the perceptual properties of objects and the lexical form class cues are integral to understanding how children come to learn new words.

Introduction

How do children come to extend words to new instances? This question is at the heart of research understanding language development partially because much of language learning presumably takes place using ostensive definition: children learn a label for one object, event, or property and are able to extend that label to new instances. The task used to study this is the novel word extension task. In this task a child is shown an exemplar and the exemplar is labeled. The child is then given other objects that match the exemplar on different dimensions and the child is asked to select the one that also has the same label. Although much of the research in this area tends to focus on how children extend novel count nouns, other grammatical classes have been studied as well.

For example, previous work has suggested that when encountering novel adjectives, children are likely to extend the novel adjective to other objects of the same material only if the objects are familiar to them. (Hall, Waxman, & Hurwitz, 1993). Thus, young children can extend the novel adjective "plush" to other objects of the same material if the plush objects are familiar to the child e.g. a shoe, but not if the objects are unfamiliar to the child, e.g. a widget, even when the children are provided with an adjectival syntactic frame. This finding has been interpreted as evidence that children are biased to expect a novel word to refer to a kind of object. By this account children should extend novel words to other objects of the same shape if the object is unfamiliar to them and children should extend novel words to objects sharing some other property when the object is familiar to them.

However, Landau, Smith, and Jones (1992) and Smith, Jones, and Landau (1992) have shown that young children can generalize novel adjectives to other objects that match in material. These results are seemingly at odds with Hall, Waxman, & Hurwitz (1993) because the objects presented to children in these studies were unfamiliar objects and thus by Hall et al's proposal children should initially interpret the novel words as referring to objects of the same shape or object kind.

In addition, other research has shown that children take the specific perceptual properties of objects into account when extending novel words. For example, several researchers (Soja, 1992; Dickinson, 1988; but see Markman & Wachel, 1988 for contradictory findings) have demonstrated that children extend novel nouns to solid objects with the same shape, but children extend novel nouns to non-solid substances with the same material as an exemplar. Further, Imai and Gentner (1997) have shown that children as young as two years of age generalize simple objects, complex objects, and substances differently.

Based on these previous results, we propose an additional constraint that may guide whether children are likely to extend a novel word to objects that match an exemplar in shape vs. objects that match in other properties. We call this the perceptual complexity hypothesis. By this hypothesis, complex objects, objects like tractors with multiple parts may be labeled by more possible words than simple objects, like ball. This may foster attention to shape if these other labels point to properties (wheels, smokestack, engine) are themselves correlated with shape. Thus, we predict that attention to shape should be a stronger pull when the objects are complex than when they are simple. This idea is supported by previous findings by Imai and Gentner (1997) that show when Japanese 2 year olds

are presented with simple objects or materials they are likely to extend by material. However, when Japanese 2 year olds are presented with complex objects they are likely to extend by shape.

We test this idea by presenting children with objects that are either perceptually simple or complex and objects that are either familiar or unfamiliar. If perceptual complexity matters for word extension we would expect children to more readily select objects that match in material when the objects are simple. However, if the perceptual complexity of the objects is unimportant for word learning we would expect to see no differences between the complex and simple objects. In Experiment 1 we present the novel words in an adjectival frame. In Experiment 2 we present the novel words in a count noun frame.

Experiment 1

In Experiment 1 we ask whether children are more likely to extend novel adjectives to objects of the same material or shape when the objects are simple or complex and familiar or unfamiliar. We do so by providing children with a novel word in an adjectival syntactic frame. Previous work has shown that when novel words are presented in a adjectival frame children may match by material kind (Hall, Waxman, & Hurwitz, 1993; Landau, Smith, & Jones, 1992; Smith, Jones, & Landau, 1992).

Method

Participants Fifty-six 4-year-olds participated. Half were male and half were female. The four year olds ranged in age from 48 to 59 months. Fourteen children (7 boys and 7 girls) were randomly assigned to each of four conditions. Children were tested individually in their preschools during normal school hours or in the laboratory.

Design Subjects were assigned to one of four conditions. In each condition the stimuli presented varied in the level of shape complexity (simple vs. complex) and familiarity (familiar or unfamiliar). Simple shapes were defined as objects that were composed of one or two parts whereas complex shapes were defined as objects that were composed of many parts. We assessed the shape complexity of the objects by asking 10 undergraduates to rate each of the 48 objects used in the experiment on shape complexity using a 5-point scale. Objects that were judged as not very complex were given a score of 1 and objects that were judged as very complex were given a score of 5. Table 1 shows the mean complexity ratings for each of the four conditions. As can be seen objects in the simple conditions were judged as less complex than objects in the complex conditions.

Table 1: Shape Complexity Ratings for the Four Conditions

Condition	Rating
Simple familiar	1.41 (.59)
Simple unfamiliar	1.53 (.61)
Complex familiar	3.72 (.97)
Complex unfamiliar	3.94 (.91)

We defined objects as familiar if they were listed on the MacArthur Communicative Development Inventory: Words and Sentences, (Fenson et al, 1994) a checklist of words known to 50% of all children by 30 months of age. However, two objects, heart and bucket/pail, were not included on the MacArthur, but pre-testing indicated these objects were known by many four-year-old children. To ensure that unfamiliar objects were truly unfamiliar to children, the unfamiliar objects were made in the laboratory and did not resemble any nameable objects.

Four triads of objects were used in each condition. In each triad there was a target object, a shape matching object and a material matching object. The shape matching object matched the target in shape and object kind but differed in material kind and related properties such as color and texture. The material matching object matched the target in material kind and related properties but differed in shape and object kind. Figure 1 shows an example of a triad from each of the four conditions. In each condition the exemplar object matched one object by material and one by shape. The four material matches were blue plush, natural wood, silver metal, and paper.

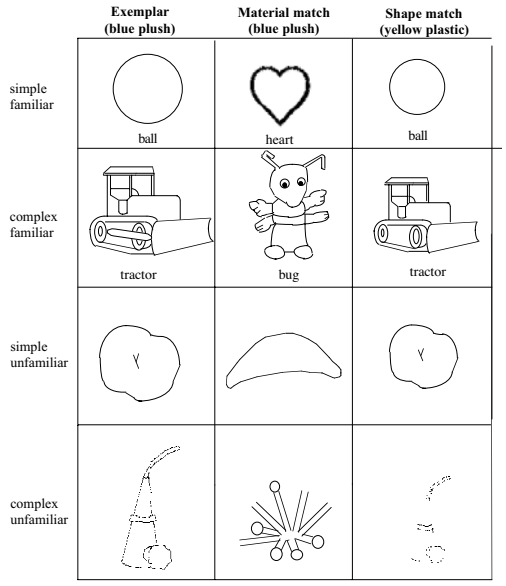


Figure 1 Examples of stimuli for the four conditions of Experiments 1 and 2.

Procedure In each condition of the adjective extension trials children were presented with an exemplar object while the experimenter labeled it with a novel adjective. For example, "See this? This is very wuggish. Can you say wuggish?" A shape match and a material match were then placed in front of the child, and the child was asked, "Can you find another one that is very wuggish?" This process was repeated for all four sets in the condition. The words blickish, fepish, wuggish, and zavish were used as the novel adjectives. The novel adjectives and the adjective syntax were taken from Hall, Waxman, and Hurwitz (1993).

Children were next presented with familiarity trials to ensure that the familiar objects were indeed familiar and the unfamiliar objects were unfamiliar to the children. Children were first "trained" to answer the familiarity trials by first presenting them with one familiar object (a shoe). Children were asked "What is this? What's this called?" Children were liberally praised for correctly labeling the shoe. Children were next presented with two unfamiliar objects and again asked "What is this? What's this called?" Children were liberally praised if they responded "I don't know." If children labeled the unfamiliar object the experimenter responded by telling the child that it was appropriate to say "I don't know" if they did not know the name of the object and children were encouraged to reply "I don't know" and were again liberally praised. Children were then randomly presented with the 12 stimulus objects from the adjective extension trials and asked of each "What is this? What's this called?" During these 12 trials children were not provided with feedback.

Results and Discussion

We first asked whether the objects we deemed as familiar were indeed familiar to children and the objects we deemed unfamiliar were indeed unfamiliar. Children responded with an appropriate label for the objects in the simple familiar condition 97% of the time on average and for objects in the complex familiar condition 91% on average. In contrast children responded that they did not know the name of the objects in the simple unfamiliar condition 89% of the time on average and for the objects in the complex unfamiliar condition 71% of the time on average. The remainder of responses in the two unfamiliar conditions involved children providing a description of the object, e.g. a green thing, an inappropriate object name, e.g. that looks like a tooth, correctly naming a piece of the object, e.g. it has a ribbon on it, or providing the novel adjective, e.g. wuggish. Thus, the results of the familiarity trials confirm that the objects in the two familiar conditions were largely familiar to children and the objects in the two unfamiliar conditions were largely unfamiliar to children.

We next examined children's performance in the adjective extension trials. Figure 2 shows the mean number of material matching selections children made in each of the four conditions. As can be seen, the number of material choices was higher in the two simple conditions than in the two complex conditions. An ANOVA conducted on the number of material choices confirmed this and revealed a main effect of complexity $F(1,52) = 25.19$, $p < .01$, but no effects of familiarity (power = .36) and no interaction (power = .21). Thus the results suggest that the shape complexity of the object, and not children's familiarity with the object, affects whether children generalize a novel adjective to a material match or to a shape match.

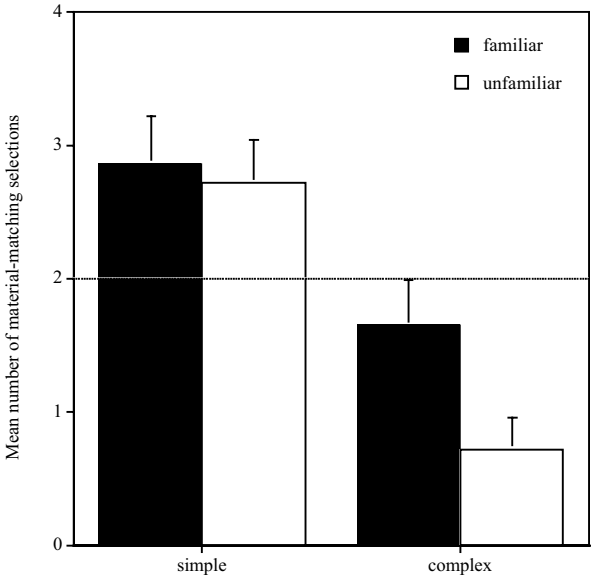


Figure 2. The number of material matching selections for the four conditions of Experiment 1.

We next compared children's selections to chance. If children responded randomly they would be expected to make material match selections in 2 of the 4 trials. The results showed that children made more material match selections than expected by chance in the simple familiar condition, $t(13) = 2.38$, $p < .05$, and in the simple unfamiliar condition , $t(13) = 2.22$, $p < .05$. Children also made less material matches than expected by chance in the complex unfamiliar condition, $t(13) = -5.26$, $p < .01$. Thus children exceed the number of material matches predicted by chance performance in the two simple conditions, but selected equal to or less material matches than predicted by chance in the two complex conditions.

Finally we asked whether knowing the object name made individual children more or less likely to match that object by material. One object included in the study made a nice test case for this question. We selected "sprinkler" as a complex familiar object in part because it was listed on the MacArthur Communicative

Developmental Inventory indicating that over 50% of all 30 month olds had produced the term. Thus we expected that 4-year-old children should be able to readily identify the object. However, only seven of the fourteen 48-60 month olds were able to appropriately identify the object as a "sprinkler" or a "sprayer". We thus asked whether correctly labeling the object coincided with more or less material choice matches. Table 2 presents the distribution of material and object matches for the 7 children who produced "sprinkler" and the 7 children who responded "I don't know" when asked what the sprinkler was. As can be seen the distributions are exactly equal suggesting that the ability to produce the basic level object name does not affect whether children are more or less likely to make a material kind selection.

Table 2: Number of material and object selections for children who did and did not produce the label "sprinkler"

	produced "sprinkler"	did not produce "sprinkler"
Object match	4	4
Material match	3	3

However, one possible explanation for our results could be that children are selecting material matches in the two simple conditions, not because they are correctly identifying the novel word as an adjective and correctly extending the word to objects that match in material in the simple object conditions, but instead that children are selecting material matches for other reasons. To test this possibility, in Experiment 2 we present children with the same sets of objects but use a count noun syntactic frame. Because count noun syntax has been shown to encourage extension to objects that match in shape, if the results of Experiment 1 are not due to idiosyncratic properties of our stimuli we would expect children in Experiment 2, who are presented with the same stimuli to make many more shape selections than material selections, regardless of the complexity of the stimuli.

Experiment 2

In Experiment 2 we again ask whether children are more likely to extend novel words to objects of the same material or shape when the objects are simple or complex and familiar or unfamiliar. We do so by providing children with a novel word in a count noun syntactic frame.

Method

Participants Fifty-six 4-year-olds participated. Half were male and half were female. The four year olds ranged in age from 4-0 to 4-11. Fourteen children (7 boys and 7 girls) were randomly assigned to each of four conditions. Children were tested individually in their preschools during normal school hours.

Stimuli and Design The stimuli and design were identical to Experiment 1

Procedure The procedure was identical to Experiment 1 with one exception. The novel word was provided to children in a count noun syntax (instead of adjective syntax). For example, "See this? This is a wug. Can you say wug?"

Results

To confirm the findings of the first experiment, we again asked whether the objects we deemed as familiar were indeed familiar to children and the objects we deemed unfamiliar were indeed unfamiliar. Children responded with an appropriate label for the objects in the simple familiar condition 97% of the time on average and for objects in the complex familiar condition 91% on average. In contrast children responded that they did not know the name of the objects in the simple unfamiliar condition 97% of the time on average and for the objects in the complex unfamiliar condition 91% of the time on average. The remainder of responses in the two unfamiliar conditions involved children providing a description of the object, e.g. a green thing, an inappropriate object name, e.g. that looks like a tooth, correctly naming a piece of the object, e.g. it has a ribbon on it, or providing the novel noun, e.g. a wug. Thus, the results of the familiarity trials confirm that the objects in the two familiar conditions were largely familiar to children and the objects in the two unfamiliar conditions were largely unfamiliar to children.

We next examined children's performance in the extension trials. Figure 3 shows the mean number of material matching selections children made in each of the four conditions. As can be seen, children generalized the novel name to the material matching object infrequently in all four conditions. An ANOVA conducted on the number of material choices revealed no main effects and no interactions. Thus, when children are provided with count noun form class cues, neither the shape complexity of the objects or children's familiarity with the basic level label of the objects affect whether children generalize a novel count noun to a material match or to a shape match. That is, children generalize by shape regardless of the particular object and its perceptual properties.

We next compared children's selections to chance. If children responded randomly they would be expected to make material match selections in 2 of the 4 trials. The results showed that children made less material match

selections than expected by chance in all conditions: the simple familiar condition, $t(13) = -6.27$, $p < .01$, the simple unfamiliar condition , $t(13) = -7.87$, $p < .01$, the complex familiar condition, $t(13) = -15.69$, $p < .01$, and the complex unfamiliar condition, $t(13) = -8.63$, $p < .01$. Thus these results confirm that children made less material matches, that is more shape matches, than expected by chance regardless of the particular object condition.

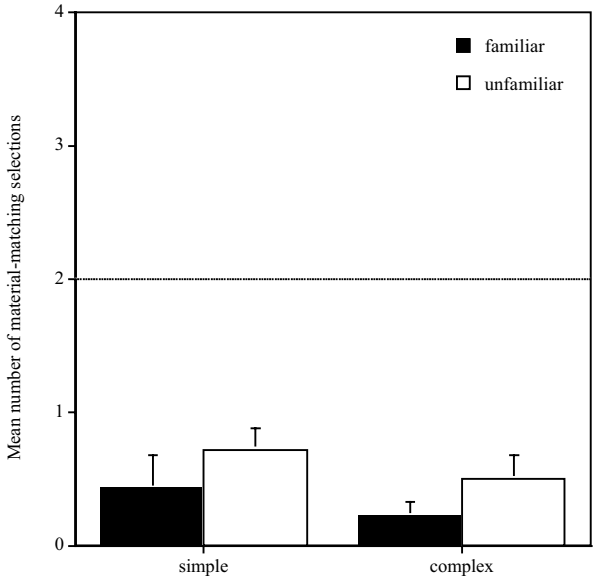


Figure 3. The number of material matching selections for the four conditions of Experiment 2.

General Discussion

Thus, the results suggest that given a count noun syntactical frame children selected objects that matched the exemplar by shape. However, given a adjectival syntactical frame children more often selected objects that matched the exemplar by material in the two simple-object conditions and shape in the two complex object conditions.

These results conflict with previous findings that children extend novel adjectives to objects of the same material only when the object kinds are familiar to the child. One possibility for this discrepancy is that the unfamiliar stimuli selected by Hall Waxman, and Hurwitz (1993) may have been inadvertently more complex than the familiar stimuli. Because Hall Waxman, and Hurwitz sought to control for taxonomic kind between the two conditions, the types of items that were unfamiliar to children may have also been slightly more complex than items that were familiar to children. For example, in the familiar condition one set of stimuli contained a cup and a spoon. In the unfamiliar condition the analogous set contained a garlic press and an apple corer. Thus the stimuli used in Hall Waxman, and Hurwitz may have inadvertently confounded familiarity with complexity.

These findings may help by providing a unifying explanation for discrepant results in the literature. One reason why some research (Hall, Waxman, & Hurwitz, 1993; Markman & Wachel, 1988) may have found that familiarity is necessary for enabling children to extend novel words by properties other than shape may have much to do with the perceptual properties of the stimuli presented to children.

Acknowledgements

This study was supported in part by a National Institute of Mental Health grant F31MH12614 to the first author

References

Dickinson, D. (1988). Learning names for materials: Factors limiting and constraining hypotheses about word meaning. *Cognitive Development, 3*, 15-35.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D., & Pethick, S. (1994). Varibility in early communictive development. *Monographs of the Society for Reserch in Child Development, 5*, whole serial number 242.

Hall, D. G., Waxman, S. R., & Hurwitz (1993). How two- and four-year-old children interpret adjectives and count nouns. *Child Development, 64*, 1651-1664.

Imai, M. & Gentner, D. (1997). A cross-linguistic study of early word meaning: Universal ontology and linguistic influence *Cognition, 62*, 169-200.

Markman, E. & Wachel, G. (1988). Children's use of mututal exclusivity to constrain the meanings of words. *Cognitive Psychology, 20*, 121-157.

Soja, N. (1992). Inferences about the meanings of nouns: The relationship between perception and syntax. *Cognitive Development, 7*, 29-45.

Decomposing Interactive Behavior

Michael J. Schoelles (mschoell@gmu.edu)
&
Wayne D. Gray (gray@gmu.edu)
George Mason University
Fairfax, VA 22030 USA

Abstract

Interactive behavior emerges from the interaction of embodied cognition with task and the artifacts designed to accomplish the task. The current study focuses on how subtle changes in interface design lead to changes in the cognition, perception, and action operations that compose interactive behavior. The Argus Prime task is explained and the nature of the modeling effort is discussed. Insights obtained by exploring differences between model and human performance in one aspect of the Argus Prime task are presented.

Introduction

The Argus Prime simulated task environment (Gray, in press) places subjects in the role of radar operators whose job it is to assess the threat value of targets on a radar display. Our goal is to determine the strategies that people use in performing the task and to study how these strategies change as a function of subtle changes in interface design. Cognitive models are built that implement these strategies at the embodiment level (Ballard, Hayhoe, Pook, & Rao, 1997). Changes in strategy that accompany changes in the interface are interpreted as due to least-effort trade offs among the cognitive, perceptual, and action elements of embodied cognition. This work has implications for interface designers of dynamic systems characterized by rapid shifts of attention and time-pressured decision making such as in air traffic controllers, emergency medical systems, and nuclear power plant systems.

Our models are written using ACT-R/PM (Byrne & Anderson, 1998) — an architecture of cognition that enables us to capture the parallelism between cognition, perception, and action. By getting the interactions right at the embodiment level (approximately one-third of a sec), we hope to reproduce process and outcomes all the way up to the scenario level (each scenario requires 12-15 min to complete).

In comparing our models to human performance, we have been alternatively pleased and disappointed. It is not uncommon for our models to match the overall performance of our human subjects (at the 12-15 min level) only to mismatch greatly at a finer level of analysis.

When part of the model misfits its part of the data, we attempt to base changes of the model on a combination of two classic approaches. First, we observe subjects and analyze action protocols of their behavior. The action protocols include response times, eye movements, and mouse movements. Second, we introduce a small change to one part of the interface. We then run the model on the two versions of Argus and compare it s predictions with empirical data collected from human subjects.

Subtle changes in interface design may result in large changes in the strategies used to perform the task. For example, in Argus Prime it is important to maximize time on unclassified targets by, in part, minimizing time spent on targets that have already been classified. Hence, a change in interface design that varies the display-based indication of a target s classification status (classified or not classified) may have a profound effect on the number and combination of cognition, perception, and action operations used to perform Argus Prime.

In this paper, we marshal both human and model data to interpret the effect of interface changes on cognitive as well as on perceptual-motor performance. We use a broad brush to describe our task, current study, and model. After discussing how well the model s performance matched overall human performance, we limit the rest of the paper to two subparts of the Argus Prime task; namely, target selection and target check. These subparts provide an example of how subtle changes in interface design can produce unexpected interactions at the embodiment level.

Argus Prime: Simulated Task Environment

Argus Prime is a complex but tractable simulated task environment. In Argus Prime the subject s task is to assess the threat value of each target in each sector of a radar screen depicted in Figure 1. The screen represents an airborne radar console with ownship at the bottom. Arcs divide the screen into four sectors; each sector is fifty miles wide. The task is dynamic since the targets have a speed and course. A session is scenario driven; that is, the initial time of appearance, range, bearing, course, speed, and altitude of each target are read from an experimenter-generated file. The

scenario can contain events that change a target s speed, course, or altitude. New targets can appear at any time during the scenario.

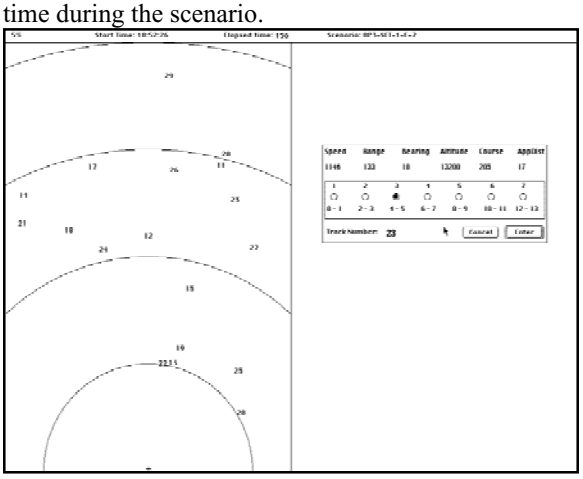


Figure 1: Argus Prime Radar Screen (left) and Information/Decision Window (right)

The subject selects (i.e., hooks) a target by moving the cursor to its icon (i.e. track number) and clicking. When a target has been hooked, an information window appears that contains the track number of the target hooked and the current value of target attributes such as speed, bearing, altitude, and course. The subject s task is to combine these values, using an algorithm that we have taught them, and to map the result onto a 7-point threat value scale (shown on the right side in Figure 1.)

Targets must be classified once for each sector that they enter. If a target leaves a sector before the subject can classify it, it is considered incorrectly classified and a score of zero is assigned.

For the versions of Argus Prime discussed here, immediate feedback was provided for each classification decision. In addition, summative feedback was provided on the percentage of targets correctly classified. (See Schoelles & Gray, in press, for more details.)

Empirical Study

This paper provides a partial report on the third study that we conducted. The results of prior studies indicated that a ubiquitous feature of the task was keeping track of which targets had been classified. In those studies there was nothing on the radar screen to indicate whether a target had been classified; that is, when a classification was made, its on-screen icon did not change (noChange). However, in both studies, if an already classified target was reselected, the target s current classification (CC) was shown in the information window (i.e., its radio button, see Figure 1, remained highlighted). We call this combination of no change to the target s on-screen icon and persistence of the classification in the information window the noChange-CC interface.

The current study manipulates the ease of retrieving status information (i.e., Is this target classified?) from the display. In addition to noChange-CC, two new interfaces are used. The noChange-noCC interface is similar to the noChange-CC in that the target s on-screen icon does not change when a classification is made. It differs from noChange-CC in that the information window contains no record as to whether a target is currently classified (i.e., once the ENTER key is pressed, the radio button is unhighlighted, see Figure 1). In contrast, for the Change interface the on-screen icon for targets that have been classified changes color. When a target is no longer classified (i.e., when it crosses a sector boundary) the icon reverts to the unclassified color.

In the first two studies subjects frequently reselected already classified targets. Their pattern of behavior suggested that for the noChange-CC interface subjects did not try to remember whether a target had already been classified. Rather, the pattern suggested that subjects simply clicked on targets until they found one that was not classified.

It was unclear in the previous studies whether this memory-less strategy (Ballard, Hayhoe, & Pelz, 1995) is adopted by choice or whether, under the conditions of the study, human cognition is incapable of retrieving target status information. This issue is tested empirically and analytically by the data and models built for the current study.

Performance on the noChange-CC interface is used as a baseline with which to compare the other conditions. We expect the noChange-noCC interface to force the memory versus memory-less issue. If subjects have no memory for having classified a target, they will be required to waste time re-computing the algorithm to reclassify already classified targets. In contrast, the Change condition provides a memory-less way to avoid classified targets and to focus on unclassified ones. Hence, subtle changes in the interface will enable different sets of strategies between the three conditions. These different strategies are expected to be differentially successful and to result in stable differences in performance.

The experiment was conducted over two sessions. In the first 2-hr session the subjects were instructed on how to do the task, did a sample scenario with the experimenter, and then did five 12-min scenarios in the noChange-CC condition. In the second 2-hr session the subjects did a 12-min practice scenario in the noChange-CC condition and then did two scenarios in each of the three conditions (noChange-CC, Change, noChange-noCC).

The Model

Our model runs under ACT-R/PM with the Eye Movements and Movements of Attention Extension (EMMA) (Salvucci, 2000). The ACT-R/PM architecture combines ACT-R s theory of cognition (Anderson & Lebiere, 1998) with modal theories of visual attention and motor movement (Kieras & Meyer, 1997). ACT-R/PM explicitly specifies timing information for all three processes as well as parallelism between them. The software architecture facilitates extensions beyond the modal theory of visual attention and motor movements.

The ACT-R/PM code executing the model runs as a separate process from Argus Prime. This process starts when the scenario starts. All communication between the model and Argus Prime is through the motor and vision module commands of ACT-R/PM.

Model Description

The recurrent task of hooking a target can be analyzed into a series of unit tasks (Card, Moran, & Newell, 1983): target selection; target check; target classification; and feedback. Each unit task has memory retrieval, visual attention, and mouse movement requirements. In ACT-R retrieval latency is a function of the activation of the memory element being retrieved. In addition, if the activation of a memory element is not above threshold, the retrieval will fail.

Movement of attention is a combination of two ACT-R/PM commands. The *Find Location* command is a pre-attentive search for a feature that returns a location to use as a parameter in the *Move Attention* command. The Move Attention command encodes a declarative memory element representing the visual object at the specified location. With the EMMA extension, a series of eye movements follows the initiation of the move attention command. The time to encode the visual object is a function of the eye movements.

Mouse movements are executed via ACT-R/PM s *Move Mouse* command. The input to this command is an object representation. The time to complete the movement is a function of Fitts Law. Mouse clicks are executed with the *Click Mouse* command. The overall operation of the model is an interleaving of productions that perform the cognitive operations of memory retrieval and goal modifications with the perceptual-motor operations of pre-attentive search, movement of attention, eye, and mouse movement.

In this paper, we focus on the target selection and target check unit tasks. In all three conditions (noChange-CC, noChange-noCC, and Change) the model begins target selection by retrieving a memory trace of the area in which it is currently searching; for example, the lower right-hand portion of the radar screen. It then pre-attentively searches for targets within this area. If a target is found, attention is moved to the feature to encode the target. (The track number is part of the encoded representation.) At that point the Move Mouse command fires and the cursor moves to and clicks on the target.

The above procedure varies slightly as a function of interface condition. In the noChange-CC and the noChange-noCC conditions, after a target is found and encoded, but before the Move Mouse command is executed, the model attempts to perform a target check by retrieving an episodic trace of a previous classification of the track number. If it retrieves this trace then it knows that the target is already classified; hence, the model will search for another target. If it cannot retrieve the trace, then the actions of moving the cursor to the target and clicking on it are performed.

In the noChange-CC condition, after clicking on a target the model will do a second target check by conducting a feature search in the information window to detect the highlighted radio button. If one is found the search for a new target will begin. Otherwise the Target Classification unit task begins. The noChange-noCC condition does not have this double-check. If it cannot retrieve a memory that the target is already classified, it will reclassify the target.

In the Change condition, targets change color after they are classified. As a consequence, the distinction between the target selection and target check unit tasks disappears. Hence, after a search area is retrieved, the pre-attentive search looks for the color feature that separates the unclassified targets (yellow) from the classified ones (blue). This strategy is purely memory-less in that no use is made of the episodic information regarding a target s prior classification. (After all targets in the retrieved area have changed color to blue, the model will do a feature search over the entire screen for yellow targets.)

Model and Subject Data Comparison

There are three limits to the model and analysis. The first is that the model was not fit to individuals. The same configuration and architectural parameters were used for all runs of the model.

Second, within an interface condition, the model uses the same strategies throughout the scenario. For example, the model uses the same target selection and target check strategies for the initial phase of the scenario when no targets had been classified as for the later stages when most targets had been classified.

Third, the base model was developed on the noChange-CC condition. In general, the way in which the model performed each unit task (i.e., target selection, target check, target classification, and feedback) was based on strategies that we observed our subjects using in the first two studies. As the unit task strategies required from 3 to 30 sec to execute, our

caveat is that in ACT-R/PM these strategies were implemented at the embodiment level using productions that required 50-100 msec to execute. Hence, the implementation of the various unit task strategies required us to make assumptions regarding memory retrieval, attention shifting, and motor movements for which we did not yet have empirical support.

In summary, at the unit task level of analysis the models implemented strategies that were cognitively plausible. For example, for the Change interface condition, pilot subjects told us that they performed target selection by looking for yellow targets. The implementation of such strategies at the embodiment level was based on our knowledge of the pre-attentive search literature, the ACT-R/PM cognitive architecture, and inspired guesses. The testing of those inspired guesses is what the current effort is all about.

Statistics Used

To compare human and model data, we report ANOVA and planned comparisons. A measure of variability for our model subjects is derived as follows. For each of 12 human subjects, a model subject was created. This model subject received the same six scenarios as the corresponding human subjects received during the second session. That is, if subject 1 did scenario 1 and 2 in the noChange-CC condition, scenario 3 and 4 in the Change condition and scenario 5, and 6 in the noChange-noCC condition, then a model subject was run on the same set of scenarios and under the same interface conditions. Hence any variability between model subjects (within condition) can be attributed to (a) unintended differences in how the 12 scenarios were designed, and (b) the randomness built into the architecture. Unlike human subjects, all model subjects in the same condition always follow the same strategies.

We confess that our second set of statistics is a willful abuse of ANOVA. The practical outcome of this is to inflate our Type I error rate; that is, the reduced variability due to model subjects should lead us to identify more differences between model subjects and human subjects than actually exist. We accept this inflated Type I error rate. The cost to our research of an inflated Type I error rate will be to cause us to spend time and attention looking for differences in strategies at the embodiment level that may not exist.

For each figure, we present the 95% confidence interval for human subjects and model subjects. The root mean square deviation (RMSD) of each human subject from his or her scenario-matched model subject is also reported.

Total Task Performance Comparison

Interface condition has a significant effect on the performance of human subjects, $F(2, 22) = 31.71$, $p < .0001$, $MSE = 33.9$. As Figure 2 shows, the Change condition does best (85.5%), followed by noChange-CC (76.4%), with noChange-noCC (66.6%) the worst. Planned comparisons show that the difference between each pair of conditions is significant at the level of significance adopted for this report ($p < .05$).

Overall, our model subjects do about as well as our human subjects (see Figure 2). However, although the main effect of model versus human is not significant, $F(1, 22) = 1.64$, $p = .21$, the interaction by interface condition is [$F(2, 44) = 7.3$, $p < .002$]. The Figure suggests that humans do slightly better than the model for noChange-CC and Change conditions but about equal to the model for noChange-noCC. As Figure 2 shows, the model makes the explicit prediction that the two noChange groups will have equal performance (noChange-CC = 67.8%; noChange-noCC = 67.5%). The significant interaction suggests that our human subjects are reacting to the interface conditions in a way that the model does not.

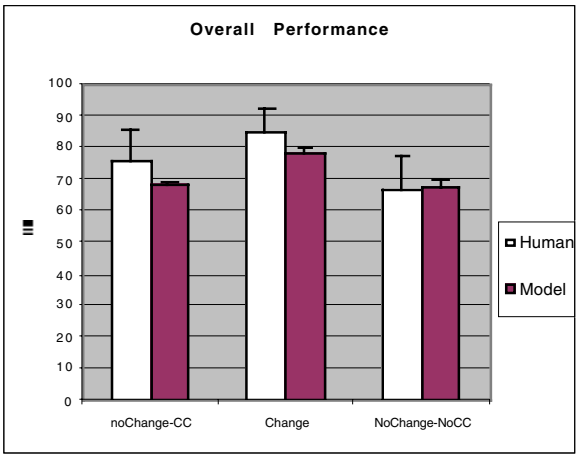


Figure 2: Overall Performance comparison of human and model. The RMSDs are 17 for noChange-CC, 13 for Change and 16 for noChange-noCC.

The variability shown by the model reflects differences between scenarios, not differences in strategies within conditions, and not differences inherent to individual subjects. Hence our model-driven approach provides us with an independent way of accessing the equivalence of the scenarios. As shown by the small confidence interval for the model subjects, our efforts to create equivalent scenarios was largely successful.

Target Selection Comparison

The three interface conditions showed significant differences in the number of *unclassified* targets

selected. The Change condition selected the most unclassified targets (68.3), followed by noChange-CC (58.3), and then by noChange-noCC (50.0). The model differences mirrored the human differences.

The more interesting comparison examines the probability of reselecting (or rehooking) an already classified target. For humans, planned comparisons show that noChange-CC rehooks the most targets (79.1) with there being no statistical difference in the number of targets rehooked by the Change (3.7) and noChange-noCC (16.3) conditions. [The overall ANOVA yields a significant effect, $F(2, 22) = 44.3$, $p < .0001$, $MSE = 441.3$.]

Comparing model performance with human performance yields a number of small surprises. Although the overall human versus model comparison is not significant ($F < 1$), we find a significant interaction of model versus human by interface condition, $F(2, 44) = 5.3$, $p < .008$, $MSE = 228.8$). Compared to humans, the model rehooks fewer targets in the noChange-CC condition, the same number of targets in the Change condition, and more targets in the noChange-noCC condition.

In both noChange conditions, prior to selecting a target the model attempts to retrieve a memory of whether that target had been classified. Only if the retrieval fails will the model rehook an already classified target. The fact that humans rehook more targets than the model in the noChange-CC condition, implies that humans rely less on memory retrieval, in this condition, than does the model. In this condition, a memory check is, in some sense, unnecessary as clicking on the target will open the information window that will clearly show whether a radio button is highlighted or not.

It may be that the cost of a perceptual-motor check is so much less than the cost of encoding and retrieving a memory that the noChange-CC condition relies on a single activity strategy, rather than one that involves dual activities (i.e., memory and perceptual-motor).

To investigate this further, the model was modified to only perform a perceptual-motor check; that is, to exclude the attempted memory retrieval. As shown in Figure 3, the model without memory selects many more targets than do human subjects. The fact that the two models bracket human performance (see also Gray & Boehm-Davis, 2000) suggest that the memory-less strategy is the preferred but not the exclusive strategy. We are currently interrogating the human data for clues as to the circumstances under subjects in the noChange-CC condition will use a memory retrieval strategy.

In contrast to noChange-CC, the perceptual-motor strategy is not available to the noChange-noCC condition. In this condition, the cost of the failure of the memory retrieval strategy is high. Reclassifying an already classified target is effortful; consuming time that would be better spent classifying an unclassified target. Hence, this greater downstream cost may lead human subjects to encode a memory trace to a higher level of activation than the model. Alternatively, it may lead them to more attempted retrievals than the model or, perhaps, to adopt a lower retrieval threshold than the model.

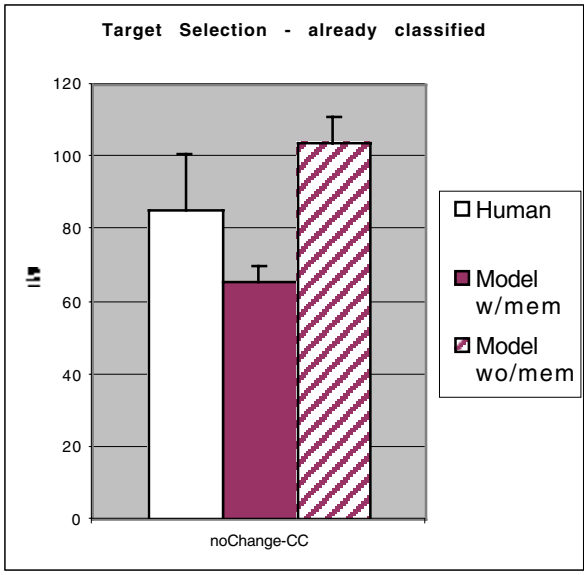


Figure 3: Target Selection for Human and Model with and without a memory retrieval. Model w/mem attempts to retrieve an episodic trace of the encoded target; only if the retrieval fails will a perceptual-motor check be performed. Model wo/mem will only perform the perceptual-motor check. The RMSDs are 42 for the human and model w/mem and 47 for the human and model wo/mem.

Our empirical data shows that subjects in the noChange-noCC condition reliably require 100 msec more than in the other conditions to classify a target. As it is not obvious why classification per se should take longer. However, the extra 100-msec is just enough time to sneak in one extra retrieval of the trace of the target just classified (Altmann & Gray, 1999), thereby increasing the success of the memory strategy for the Target Check unit task. Models incorporating this 100 msec of extra strengthening are being built and will be tested to determine if this strengthening suffices to produce the increment in performance shown by humans over the current model.

Discussion

Performance of the model subjects can be viewed as the embodiment of our theory of human performance. Comparisons that yield a significant main effect of model versus human signal places where our theory of human performance breaks down. Comparisons that yield a significant interaction of model versus human signal places in which our understanding of how

interface design influences interactive behavior are deficient. With this as our perspective, what does the performance of our model subjects tell us about our understanding of human interactive behavior?

A message that comes through loud and strong is that if our goal is to understand cognitive processes and not simply to predict performance outcomes, then obtaining good fits to an overall performance measure, such as Total Task Performance, can be misleading. The fit between model and human on overall performance can mask large and important differences in unit task performance.

On the first two unit tasks, Target Selection and Target Check, neither the main effect nor interaction of model versus human was significant for number of unclassified targets selected (first hook). This excellent fit of model to data broke down when we examined the number of times a target was rehooked. In this case the interaction indicated much more rehooking for noChange-CC than expected and less rehooking for noChange-noCC than expected. This interaction could be explained if the noChange-CC condition relied more on a perceptual-motor strategy and less on memory than did the model. Similarly, the noChange-noCC condition may be encoding the episodic trace of already hooked targets more highly than we had anticipated.

Conclusions

The goal of our research effort is to understand how subtle changes in interface design may lead to large changes in overall performance. As interactive behavior emerges from the interaction of embodied cognition with task and the artifacts designed to accomplish the task, an explanation of performance changes requires a consideration of the fine details of this interaction. In this article we have focused on one type of change and its effect on one part of task performance.

Although the fit of our model to overall performance was good, examining the fit of the model at the unit task level revealed important mismatches. For the Target Selection and Target Check unit tasks, the initial selection of unclassified targets was well fit by the model but the rehooks were not. Analyses of the model and the ways in which it matched and mismatched the data suggested three distinct target checking strategies that varied in their reliance on perceptual-motor versus memory operations.

Acknowledgments

This work was supported by Air Force Office of Scientific Research Grant # F49620-97-1-0353. We thank the many members of the Argus Group who have contributed to the Argus Prime studies: Erik M. Altmann, Deborah A. Boehm-Davis, Jeni Paluska, and Ryan Snead.

References

Altmann, E. M., & Gray, W. D. (1999). Serial attention as strategic memory. In M. Hahn, & S. C. Stoness (Eds.), *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society* (pp. 25-30). Hillsdale, NJ: Erlbaum.

Anderson, J. R., & Lebiere, C. (Eds.). (1998). *Atomic components of thought*. Hillsdale, NJ: Erlbaum.

Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience, 7*(1), 66-80.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences, 20*(4), 723-742.

Byrne, M. D., & Anderson, J. R. (1998). Perception and action. In J. R. Anderson, & C. Lebi re (Eds.), *The atomic components of thought* (pp. 167-200). Hillsdale, NJ: Erlbaum.

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gray, W. D. (in press). Simulated task environments: The role of high-fidelity simulations, scaled worlds, synthetic environments, and microworlds in basic and applied cognitive research. *Special Joint Issue of Cognitive Science Quarterly and Kognitionswissenschaft*.

Gray, W. D., & Boehm-Davis, D. A. (2000). Milliseconds Matter: An introduction to microstrategies and to their use in describing and predicting interactive behavior. *Journal of Experimental Psychology: Applied, 6*(4), 322-335.

Kieras, D. E., & Meyer, D. E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction, 12*(4), 391-438.

Salvucci, D. D. (2000). A model of eye movements and visual attention. In N. Taatgen, & J. Aasman (Eds.), *Proceedings of the Third International Conference on Cognitive Modeling* (pp. 252-259). Veenendal, The Netherlands: Universal Press.

Schoelles, M. J., & Gray, W. D. (in press). Argus: A suite of tools for research in complex cognition. *Behavior Research Methods, Instruments, & Computers*.

The Influence of Causal Interpretation on Memory for System States

Wolfgang Schoppek (wolfgang.schoppek@uni-bayreuth.de)
Department of Psychology, University of Bayreuth
D-95440 Bayreuth, Germany

Abstract

This paper reports an experiment that investigated the influence of causal interpretation on acquisition and use of two knowledge types about a static system: I-O knowledge (instances of system states) and structural knowledge (knowledge about causal relations within the system). One group of subjects saw system states without being informed about the causal nature of the material. Another group saw the same states as switches and lamps. It is assumed that the group without causal interpretation can only acquire I-O knowledge. If I-O knowledge is the predominant type when dealing with small systems, then there should be no group differences in a recognition task. Actually, the group with causal interpretation discriminates much better between targets and distractors, but with longer RTs. This is interpreted in terms of structural knowledge acquired by the group with causal interpretation, which was used to reconstruct system states in cases of doubt. Results of a task where subjects had to judge single causal relations support that interpretation, but also indicate that the knowledge about effects is probably not represented in an explicit, symbolic form. An ACT-R model that uses associations between events as a subsymbolic form of structural knowledge reproduces the data well. Thus, data and model support the significance of I-O knowledge but also shed some light on the role and the development of structural knowledge.

One central question in the psychological research on complex dynamic systems refers to the knowledge that is used for controlling a system. One important aspect of that question refers to the content of the acquired knowledge. Subjects may acquire structural knowledge, defined as general knowledge about the variables of a system and their causal relations. They may as well acquire input-output knowledge (I-O knowledge), which represents instances of input values and the corresponding output values.

There is evidence for the influence of both types of knowledge on performance in system control, but currently many authors emphasize the role of I-O knowledge, particularly when dealing with small systems like the "Sugar Factory" (a dynamic system with one input and one output variable, connected by a linear equation; Berry & Broadbent, 1988). Computational models developed on the basis of Logan's Instance Theory (Dienes & Fahey, 1995) or ACT-R (Lebiere, Wallach & Taatgen, 1998) demonstrate the sufficiency of I-O knowledge for the control of the "Sugar Factory". The strategy of relying on I-O knowledge seems to be preferred by most subjects, even in the control of more complex systems. However, in systems of at least six variables, high performance is usually associated with structural knowledge (Funke, 1993; Vollmeyer, Burns & Holyoak, 1995).

A second aspect of the question as to what knowledge is used in system control refers to its status as explicit or implicit knowledge. In an experiment with the "Sugar Factory", Dienes and Fahey (1998) found stochastic independence between the solution of studied control problems and the recognition of the same situations as studied. The authors concluded that memory for the situations was implicit. This result extends the common finding of dissociations between recognition and completion tasks (e.g. Tulving & Hayman, 1993) to the domain of system control.

In the present paper these questions were investigated by using stimuli that can be either interpreted as states of a system or simply as spatial patterns. The rationale of the experiment is that learning of instances does not depend on the causal interpretation of stimuli. Consequently, if knowledge about instances (I-O knowledge) is the main knowledge type learned, there should be no effect of causal interpretation on recognition of system states. On the other hand, if structural knowledge is learned additionally, then causal interpretation should have positive effects, particularly in a causal judgment task.

The assumptions about the two knowledge types are explicated with a computational model based on the ACT-R theory (Anderson & Lebiere, 1998). The model reproduces the results of the experiment quite well, and can be considered being an explanation for the stochastic independence between completion and recognition tasks.

Experiment

The significance of I-O knowledge and structural knowledge was studied with a system consisting of four lamps operated by four switches. Figure 1 shows a screenshot with the effects of the switches mapped (the arrows were not visible for the subjects). Each switch

affects one or two lamps. Two of the effects are negative, which means that the corresponding lamp is switched off when the switch is turned on.

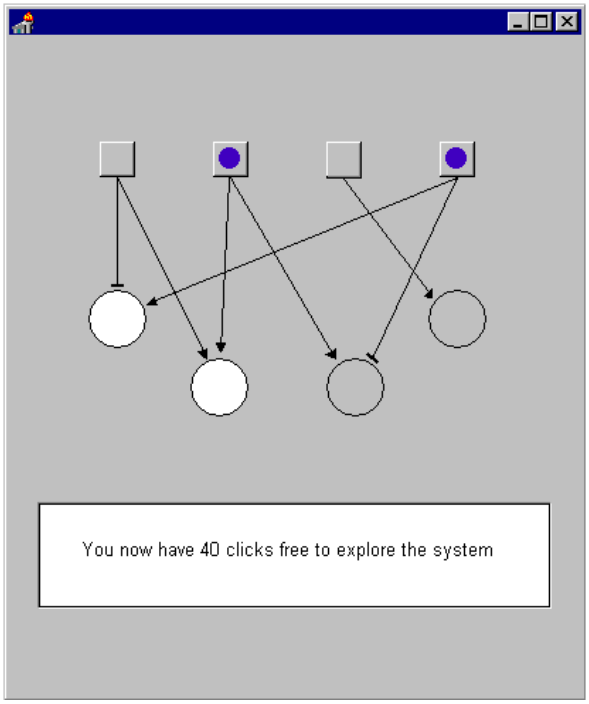


Figure 1: The system used in the experiment. The arrows were not visible for the subjects; ↓: on relation, ⊥: off relation.

Two tasks were used, each more sensitive to a different type of knowledge: A recognition task - easiest to be done with I-O knowledge, and a causal judgment task - easiest to be done with structural knowledge. Additionally, a pattern completion task was administered, which is not expected to be particularly sensitive to one knowledge type.

In the speeded recognition task subjects saw ten possible and ten impossible system states two times each, and had to decide if they had seen the state in the learning phase or not. The items of the speeded judgment task were pictures of the switches and lamps with one switch and one lamp highlighted. Subjects had to decide if there was a causal relation between the highlighted elements. The 16 possible combinations were shown twice. In the completion task subjects were shown eleven arrays of switches or lamps and asked to complete the missing parts, i.e. complete the lamps when switches were shown and vice versa.

Two factors were varied between subjects: (1) the possibility to interprete the pictures of system states shown in the learning phase as causal, and (2) the subject's activity, i.e. if the system states were either observed, or produced by operating the switches. I will focus on the effects of the first factor (that were the strongest ones, anyway), and report the data of the two groups who observed the system states in the learning phase, either with causal interpretation (ci), or without causal interpretation (nci). Each of the groups consisted of 12 subjects.

Other factors were varied within subjects: (1) the number of presentations of each state in the learning phase (1-2 presentations vs. 3-5 presentations), and (2) the number of switches that were "on" in each item of the recognition task (1 switch on vs. 3-4 switches on).

The experiment started with a learning phase where subjects saw 40 system states in intervals of four seconds. Each possible state of the system was shown at least once. The group without causal interpretation (nci) was told that they would see spatial patterns, which they should memorize. The group with causal interpretation (ci) was informed that the patterns were states of a system of switches and lamps.

Three minutes after completion of the learning phase the recognition task was administered followed by another 25 system states. Next, subjects worked on the completion task. Then the subjects of the group without causal interpretation were debriefed about the causal nature of the stimulus material. After that the judgment task was provided, followed by two other tasks that are not reported here.

Given the assumption that knowledge about the system is primarily stored as specific instances, the factor "causal interpretation" should have no effects on performance in the recognition task. If, however, subjects acquire structural knowledge - which is expected only in the group with causal interpretation - that group should outperform the nci group, particularly in the judgment task.

As a measure of performance in the recognition and judgment tasks, discrimination indices P_r were calculated according to the Two-High-Threshold-Model (Snodgrass & Corvin, 1988). A discrimination index of 1 indicates perfect discrimination; a value of 0 indicates random performance.

Table 1: Discrimination indices for two tasks

	ci	nci
Recognition	M = 0.48 s = 0.23	M = 0.30 s = 0.22
causaljudgment	M = 0.55 s = 0.18	M = 0.17 s = 0.23

Table 1 shows means and standard deviations of these indices. In both tasks the group with causal interpretation is significantly better ($F_{1,22} = 10.76$, $p < .01$), and there is an interaction between task and group ($F_{1,22} = 7.26$, $p < .05$). The ci group is better at judging causal relations than at recognition; for the ci group the reverse is true. Latencies for hits are longer in the group

with causal interpretation (ci: 2250 ms, nci: 1493 ms). The fact that the variance is also significantly higher in the ci group points to the use of different strategies: If a system state could not be retrieved in the recognition task, subjects of the ci group might have tried to reconstruct the state by using knowledge about the effects of the switches. That would mean that subjects used both, I-O knowledge and structural knowledge.

This interpretation is supported by the effects of the within-subjects factors on recognition performance (Figure 2, left panel). If the reconstruction hypothesis is true, then there should be an effect of the number of switches on in the ci group, because the reconstruction process is harder the more switches have to be considered. Actually, a significant interaction between group and number of switches on was found in the proportion of hits ($F_{1,22} = 6.13$, $p < .05$). States with three or four switches in on position are particularly badly recognized by the subjects of the ci group. On the other hand, in the group without causal interpretation the influence of number of presentations is higher (interaction marginally significant: $F_{1,22} = 3.63$, $p = .07$). All this supports the assumption that the group with causal interpretation used I-O knowledge and structural knowledge in both tasks.

Further inferences about the application of one vs. two knowledge types can be drawn from contingency analyses between the tasks. Since the mapping between the items of the recognition task and the items of the judgment task is ambiguous, I calculated contingencies between recognition and completion.

If there is only one (explicit) knowledge type, items that were completed correctly should also be recognized as studied. For two knowledge types, the contingency prediction is less clear. If each task is solved with different knowledge, stochastic independence between the tasks should be the consequence.

The items of the completion task were entered into contingency tables depending on their solution and their recognition (e.g. Item 1 was solved correctly and not recognized as studied). The entries were summed over all subjects of each condition and over all items. Empirical contingencies, measured by Δp, were compared with maximum contingencies that can result with the given marginal distributions[1]. In the ci group the empirical contingency between recognition and correct completion is 0.17. This is considerably lower than the maximum of 0.65. In the nci group the contingency is 0.41, which is much closer to the maximum of 0.53 in that group. Thus, in the group with causal interpretation the solution of completion items does not depend on correct recognition of these items as studied, whereas in the group without causal interpretation a moderate degree of dependency was found between the two tasks. Again, the results are compatible with the assumption that the nci group used only one type of knowledge, whereas the ci group used two types.

[1] The maximum possible memory dependence as suggested by Ostergaard (1992) could not be calculated because only studied items were used in the completion task.

Discussion

Overall, the results support the assumption that causal interpretation enabled subjects to gain an additional type of knowledge. This raises the question about the nature of that knowledge. In the introduction I hypothesized that it should be structural knowledge. But there is one result that is problematic for this conclusion: Since structural knowledge is ideal for solving the judgment task it is surprising that the mean latency for hits is as long as 2234 ms (see also Schoppek, 1998 for similar results). If subjects tried to retrieve structural knowledge right away, the latency should be much shorter. A possible explanation is that most subjects try to use I-O knowledge first and use knowledge about effects only after retrieval of relevant I-O knowledge fails. The reason for that might be that knowledge about causal relations is not represented explicitly in symbolic form, but rather in form of associations between events. In the ACT-R theory (Anderson & Lebiere, 1998), associations between declarative memory elements and their baselevel activations are described as the subsymbolic level of declarative memory. This level is implicit in the sense that it affects symbolic processing (e.g. retrieval) without being directly accessible. In the next section I describe a computational model that uses the distinction between symbolic and subsymbolic level to explain the effects of causal interpretation.

ACT-R Model

In order to test how the above interpretation can reproduce the data, I developed an ACT-R model that simulates the learning phase, the recognition task, and the judgment task. There are two versions of the model. One of them entails additional production rules for modeling causal interpretation. These rules reconstruct a system state when no relevant memory representation of the state can be retrieved. The state is reconstructed on the basis of associations between events.

In the learning phase a new declarative element (called chunk in ACT-R) is created for each system state and pushed on the goal stack. After processing the goal it represents a system state with its slots holding the arrays of switches and lamps. These state chunks are the basic units of I-O knowledge. Also in each cycle, a change-image is created as a subgoal, representing the changes between the previous and the current system state. Most of the change-images are not strong enough to be retrieved later on, but during goal elaboration associative weights are learned between

switch- and lamp-events (e.g. between the events "Switch A turned on" and "Lamp 1 turns dark"). Afterwards these associations are used to reconstruct system states in the condition with causal interpretation. No structural knowledge is explicitly induced, because otherwise the model would predict much shorter response times in the judgment task.

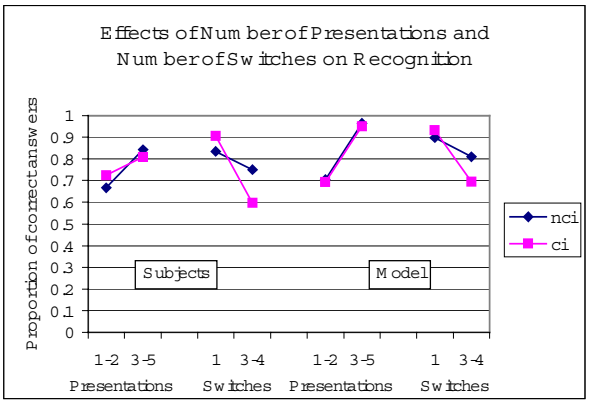


Figure 2: Experimental (left) and model (right) results

In the recognition task both model versions try to retrieve an instance similar to the probe. The constraints for retrieval are either "retrieve a chunk that has the probe's combination of switches in its switches slot" (retrieval by input), or "retrieve a chunk that has the probe's combination of lamps in its lamps slot" (retrieval by output). The model has a bias towards using the tactic of retrieval by input. Partial matching is turned on, which means that not only perfectly matching instances can be retrieved, but also instances that are similar to the retrieval constraints. If retrieval fails, the version without causal interpretation guesses, the other version starts the reconstruction process. Reconstruction is based on the lamp-events that are most strongly activated by the switch-events shown in the probe ("switch on"). The probability of false reconstructions rises with the number of switches that are on - an effect that explains the bad recognition performance under condition ci & 3-4 switches.

I simulated two samples with 24 cases each[2]. Some results are shown in the right panel of Figure 2. In both simulated between subject conditions recognition performance depends more on the number of presentations as compared to the real subjects. But the interaction between number of switches and causal interpretation is well reproduced by the model. In general, the model overestimates recognition performance. This effect is mainly due to the excellent recognition of the frequently shown system states. Latencies for hits are very close to the data: 2314 ms in the simulated ci group and 1541 ms in the simulated nci group (note that the latency factor was fitted for the nci group only).

After fitting parameters for the recognition task, the model was extended with a few production rules to solve the causal judgment task. In that task the model tries to retrieve a diagnostic instance appropriate to confirm the causal relation. For example, when the item requires judging the causal relation between Switch A and Lamp 3, the model tries to retrieve a chunk that represents the system state with Switch A as the only switch on. Assume the model retrieves the appropriate state (Switch A on, Lamp 2 on), it will produce the answer "no". If no diagnostic state can be retrieved, the model reconstructs the state in the same way as in the recognition task.

In the simulation with this part of the model, I assumed that the judgment task was done right after the learning phase. Recall that the groups of subjects that have been discussed so far did the judgment task later in the experiment. Therefore, the simulation results were compared to a group of subjects (N=12) who did the judgment task in the first place. That group was informed about the causal interpretation of the stimuli.

The model matches the subjects' data quite close without fitting any parameters (Figure 3). Mean latencies for hits were 2305 and 2234 ms in the model's and subjects' data, respectively.

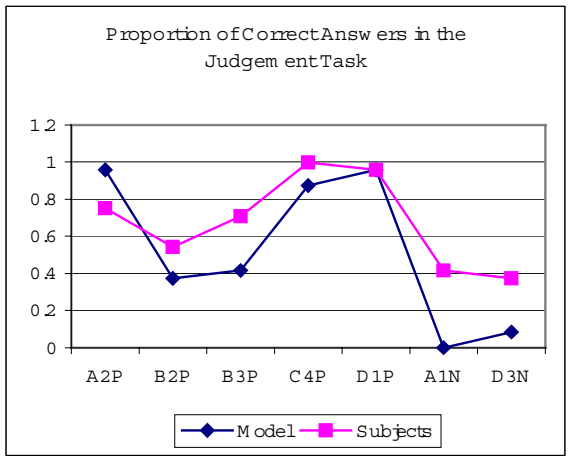


Figure 3: Proportions of correct answers in the judgment task. A2P through D1P are the five "on" relations of the system, A1N and D3N the two "off" relations. (A2P: Switch A – Lamp 2 - positive)

General Discussion

Model and data support the view that I-O knowledge is the primary type of knowledge used when dealing with a small system. But longer latencies, together with

[2] Parameter values were as follows: partial matching=on, mismatch penalty=2.5, baselevel learning=0.5, retrieval threshold=0.75, parameter learning=off, associative learning=3.0, activation noise s=0.5, expected gain noise s=0.5, latency factor=2.5. The source code of the model is available at www.uni-bayreuth.de/departments/psychologie/cogsci01.html

better recognition in the group with causal interpretation point to the use of an additional type of knowledge. It has been modeled as subsymbolic associations between events, used to reconstruct a mental image of the system state in question.

The results of the group with causal interpretation parallel the findings of Dienes & Fahey (1998), but the interpretations are slightly different. Dienes and Fahey assume that subjects learn a lookup-table of system states and conclude from their data that this table is stored in implicit memory. The lookup-table is similar to I-O knowledge. The difference is that in the present conception I-O knowledge is always explicit, and a second type of knowledge is assumed – subsymbolic associations between events. In this interpretation it is the subsymbolic knowledge that would be considered implicit.

Applying the distinction between symbolic I-O knowledge and subsymbolic associations between events to the "Sugar Factory" could explain the results of Dienes & Fahey (1998). If subjects used I-O knowledge about past situations in the recognition task and associations between events in the control problems, stochastic independence between the two tasks could be the consequence. The explanatory potential of the subsymbolic level of ACT-R for implicit memory phenomena has also been demonstrated by Taatgen (1999) with a model of word recognition and completion. In his model it is the dynamics of baselevel learning rather than associative learning that accounts for dissociations.

The present research yielded effects that are similar to those known from other paradigms. It is a common finding that providing additional information about stimuli enhances memory or other kind of performance, e.g. in classification learning (Nosofsky, Clark, & Shin, 1989), Schema acquisition (Ahn, Brewer, & Mooney, 1992), or text comprehension (Bransford & Johnson, 1973; Kintsch & van Dijk, 1978). Also the finding that most subjects spontaneously rather use I-O knowledge or knowledge about specific instances than using structural knowledge or rule knowledge has parallels in these paradigms. Nosofsky et al. (1989) found that even simple rules defining a concept were only used when subjects were explicitly told to do so. Ahn et al.'s (1992) subjects used the experimentally provided background knowledge only when they were engaged in tasks requiring the active use of that knowledge.

An important question is at what point in the whole process the causal interpretation effect arises. The present model assumes that the associations between events are learned incidentally in both conditions, and the effect occurs during recall, when only the ci subjects use this knowledge. This assumption shall be tested in future experiments.

The next step in this research is modeling the completion task to test if the model really predicts the effect of causal interpretation on the contingency between recognition and completion tasks. Further research is also necessary to explore if the effects of causal interpretation can be generalized to similar tasks. If the effects can be confirmed, the model provides an interesting basis for a more general theory about implicit memory phenomena.

References

Ahn, W., Brewer, W. F., & Mooney, R. J. (1992). Schema acquisition from a single example. Journal of Experimental Psychology: Learning, Memory, and Cognition, 18, 391-412.

Anderson J.R., & Lebiere C. (1998). The atomic components of thought. Mahwah, NJ: Lawrence Erlbaum Associates.

Berry D.C., & Broadbent D.E. (1988). Interactive tasks and the implicit-explicit distinction. British Journal of Psychology, 79, 251-272.

Bransford, J.D. & Johnson, M.K. (1973). Considerations of some problems of comprehension. W.G. Chase (Ed.), Visual information processing. Orlando, FL: Academic Press.

Dienes Z., & Fahey R. (1995). Role of specific instances in controlling a dynamic system. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 848-862.

Dienes Z., & Fahey R. (1998). The role of implicit memory in controlling a dynamic system. The Quarterly Journal of Experimental Psychology, 51A, 593-614.

Funke J. (1993). Microworlds based on linear equation systems: a new approach to complex problem solving and experimental results. In G. Strube & K.F. Wender (Eds.), The cognitive psychology of knowledge, pp. 313-330. Amsterdam: North-Holland.

Kintsch, W. & Dijk, T.A.van (1978). Toward a model of text comprehension and reproduction. Psychological Review, 85, 363 -394.

Lebiere C., Wallach D., & Taatgen N. (1998). Implicit and explicit learning in ACT-R. In F.E. Ritter & R. M. Young (Eds.), Proceedings of the Second European Conference on Cognitive Modelling (ECCM -98), pp. 183-189. Nottingham: Nottingham University Press.

Marescaux P.-J., Luc F., & Karnas G. (1989). Modes d´apprentissage selectif et nonselectif et conaissances acquises au controle d´un processus: Evaluation d´un modele simule. Cahiers de Psychologie Cognitive, 9, 239-264.

Nosofsky, R.M., Clark, S.E. & Shin, H.J. (1989). Rules and exemplars in categorization, identification, and recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 282 -304.

Ostergaard, A. L. (1992). A method for judging measures of stochastic dependence: Further comments on the current controversy. Journal of Experimental

Psychology: Learning, Memory, and Cognition, 18, 413-420.

Schoppek W. (1998). Modeling causal induction. Paper presented at the Fifth Annual ACT-R Workshop 1998, Carnegie Mellon University, Pittsburgh, PA. (http://act.psy.cmu.edu/ACT/ftp/workshop/Workshop-98/Schoppek/quick_index.html)

Snodgrass, J. G. & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. Journal of Experimental Psychology: General, 117, 34-50.

Taatgen, N. (1999). Learning without limits. Groningen, The Netherlands: Universal Press of the Rijksuniversiteit Groningen.

Tulving, E., & Hayman, C.G. (1993). Stochastic independence in the recognition/identification paradigm. European Journal of Cognitive Psychology, 5, 353-373.

Vollmeyer R., Burns B.D., & Holyoak K.J. (1996). The impact of goal specificity on strategy use and the acquisition of problem structure. Cognitive Science, 20, 75-100.

Metarepresentation in Philosophy and Psychology

Sam Scott (sscott@ccs.carleton.ca)
Department of Cognitive Science
Carleton University, Ottawa, ON K1S 5B6

Abstract

This paper brings together two definitions of metarepresentation: Dennett's notion of metarepresentation as second-order representation, and an alternative definition of metarepresentation found in the work of Leslie, Frith, and Baron-Cohen on autistic children. I show that the two definitions are not in any way compatible with one another, and that the assumption that they *are* compatible can lead to confusion about the nature of higher cognition. I illustrate this potential for confusion through the analysis of some claims made in a paper by Whiten and Byrne on primate cognition.

Representation

I will use the term "representation" to mean *mental representation* as defined in Von Eckardt's (1999) MITECS entry. Her definition of mental representation is (I hope) sufficiently broad and uncontroversial to be acceptable to most of the various competing currents in cognitive science. According to Von Eckardt, a (mental) representation has four important aspects: "(1) it is realized by a representation bearer; (2) it has content or represents one or more objects; (3) its representation relations are somehow 'grounded'; (4) it can be interpreted by (will serve as a representation for) some interpreter." (p. 527) Points (1) and (4) in the above establish that a (mental) representation requires a subject that both bears and can interpret the representation.

Point (2) establishes what the representation can be about. The point about representing one or more objects is fairly clear, but the point about "having content" needs some unpacking. Fortunately, Von Eckardt does that unpacking for us. A (mental) representation is something that can stand for "concrete objects, sets, properties, events, and states of affairs in this world, in possible worlds, and in fictional worlds as well as abstract objects such as universals and numbers; that can represent both an object (in and of itself) and an aspect of that object (or both extension and intension); and that can represent both correctly and incorrectly." (p. 527) Von Eckardt's list is probably not exhaustive, but it does cover the ability of cognitive systems to "think about" objects in the world, counterfactual situations, and propositions and predicates, all under the umbrella term *representation*.

The only point that remains undeveloped in Von Eckardt is point (3), which states that relations must be "grounded". I take that to mean simply that there must be an external referent of some kind for any representation, although this "external" referent may only exist in a possible or fictional world.

Metarepresentation

The prefix "meta" can mean a number of different things in different contexts (e.g. "metaphysics", "metaphilosophy", "metamorphosis" to name but a few) but the usual sense attributed by philosophers is that a metarepresentation is a higher-order representation of some kind. That is, a metarepresentation is a representation of a representation. Following Dennett (1998), it stands to reason that if a representation exists as an object in the world, then it too can be represented. Dennett's examples of metarepresentation tend to be of a hybrid nature. For instance a drawing on a piece of paper is a type of non-mental representation, which is represented in the mind of the person viewing it. The mental representation is of the drawing, but since the drawing is itself a representation, the viewer has a (mental) metarepresentation of whatever it is that the drawing represents.

Despite the drawing being an "external" rather than a mental representation it does share many of the properties of the latter. Following Von Eckardt as quoted above, the drawing: (1) has a representation bearer (the paper); (2) has content (whatever the drawing represents); (3) has a referent; and (4) can be interpreted by some interpreter. Most of Dennett's examples are to do with hybrid metarepresentation – mental representations of external representations. An interesting question is whether hybrid metarepresentation is the same sort of thing as purely mental metarepresentation. Some would say no, arguing that in the hybrid case, there is a difference of content – the external and the mental represent in different ways, therefore a representation of an external representation has a different type of content from a representation of a mental representation. Dennett would not want to take this approach, opting for an intentional stance in which he could avoid discussing matters such as internal content – things represent if we can sensibly treat them as representing, regardless of whether the representation has any further degree of reality. For the

current discussion, I would like to leave aside issues of content and the differences between hybrid and purely mental metarepresentation. In any case, this paper discusses purely mental metarepresentation almost exclusively. I hope I can safely take from Dennett the intuitively satisfying notion that the definition of "metarepresentation" corresponds roughly to the definition of "higher-order representation".

In addition to being intuitive, the "higher-order" definition of metarepresentation is also the one that has seen most use in philosophical circles. Unfortunately, in a large part of the psychological literature, it is not clear that this is the definition that is in use. In what follows, I will show that the so-called "metarepresentational conjecture" that is postulated to explain certain aspects of autistic behavior is making use of a very specific and technical definition of the word "metarepresentation". Of course this fact on its own should be neither surprising nor cause for alarm – any group of scientists should always feel free to redefine terms in technical ways that suit their needs. But unfortunately, the different definitions have lead to confusion even within the psychological literature. Before moving on to this literature, however, it is worth spending some time sorting out potential confusions lurking within the definitions articulated above.

What Metarepresentation is NOT

First of all, a representation can *contain* other representations without being a metarepresentation. For instance, consider the representations that might be necessary to entertain the thought corresponding to the following proposition:

(1) Mélissa's dog is dead

At the very least, we need a representation of Mélissa's dog. We will also need a representation of some one-place predicate DEAD. Finally, it is possible that we would also need a representation of the saturated predicate DEAD(Mélissa's dog). Depending on your personal biases (i.e. connectionist or classical), you may therefore want to assert that understanding sentence (1) requires a representation of Mélissa's dog which is *contained within* a representation of the predicate DEAD(Mélissa's dog). If so, this is not the same thing as the representation of DEAD(Mélissa's dog) being (even partially) a metarepresentation *of* Mélissa's dog. That is, there is nothing necessarily metarepresentational going on in this situation.

So far so good, but the next assertion may be more controversial. Second-order beliefs and desires do not necessarily require metarepresentations either. To see why, consider the following first-order belief:

(2) Mélissa BELIEVES that her dog is dead

This first-order belief requires Mélissa to have a mental representation of the proposition "my dog is dead" and believe that the proposition is true (perhaps it is marked as "true" in her mental database, or perhaps she has it in her "belief box", or whatever). The important observation here is that Mélissa need not be aware of her belief. If she were, she would require a representation of it, but to simply hold the belief, no such special mental machinery is required. She need not think to herself "I believe my dog is dead" in order to believe her dog is dead. She just needs to believe that her dog is dead. Thus "believe" is a definitional label we apply to any state of affairs in which someone holds a proposition to be true. This is why we can speak of animals having beliefs, even if we are not comfortable with the notion that they may be aware of them.

Now consider the following second order belief:

(3) Anne BELIEVES that Mélissa BELIEVES that her dog is dead

What kinds of representations do we need to ascribe to Anne in this case? First, she needs the representation of Mélissa's dog, the predicate DEAD, and so on. What she doesn't need is a representation of *Mélissa's representation of* her dog, the predicate DEAD, and so on. That is, she doesn't need a second-order representation of any of these things. She can get by with her own first-order representations. But it would appear that Anne also needs to have a representation of Mélissa's BELIEF. That is to say, she needs a representation of Mélissa's mental state of believing in a way that Mélissa does not. She must be aware of Mélissa's BELIEF, while Mélissa need not be. If we consider Mélissa's mental state of believing to be an object in the world, then this mental state must be represented somehow in Anne's belief. The question of whether we need a metarepresentation here hinges on whether Mélissa's belief state counts as a representation. But as I pointed out above, for Mélissa to simply *have* the first order belief, no first order representation of belief is required. Since neither Mélissa nor Anne has any particular need of belief representation in order to be a believer, Anne's representation of Mélissa's belief need not be second-order.

So what *does* Anne require in order to hold belief (3) above? It would seem that certain processing requirements are necessary to be able to form such complex thoughts. (Recall that what she actually believes corresponds to sentence (2) above, and not to sentence (3).) First of all, Anne must be able to perform some kind of *propositional embedding*. She needs to be able to represent Mélissa's belief as a proposition with two arguments: a representation of Mélissa, and a representation of the proposition that she believes.

Furthermore, Anne needs to be capable of dealing with *referential opacity*. She must be able to remain agnostic about the truth-value of the embedded proposition ("Mélissa's dog is dead") and recognize that it has no effect on the truth-value of the belief proposition.

"Metarepresentation" in Autism Research

A particular definition of "metarepresentation" has played a very important role in research on Autism, where researchers have proposed the existence of a metarepresentational module to explain some of the deficits that autistic people exhibit. Alan Leslie, along with Simon Baron-Cohen and Uta Frith, are the principal proponents of metarepresentational modules in the psychological literature (Leslie, 1991; Baron-Cohen, 1991). Leslie in particular has put forward the *metarepresentational conjecture*: "Autistic children are impaired and/or delayed in their capacity to form and/or process metarepresentations. This impairs (/delays) their capacity to acquire a theory of mind." (Leslie, 1991, p. 73) Before dissecting what Leslie means by "metarepresentation", let's take a quick look at the evidence on which this statement is founded.

The three most classic experiments on autistic children are the picture sequencing task, the Sally/Anne task and the Smarties task, all of which reveal a selective deficit in autistic children in understanding false beliefs. For space reasons, I discuss only the Sally/Anne task (see Leslie, 1991 for the others). In this experiment dolls are used to act out a scenario in which Sally hides a marble in a basket and leaves the room. While she is gone, Anne enters and transfers the marble to a box. Sally returns, and the children are asked, "Where will Sally look for her marble?" Autistic children consistently make the incorrect prediction that Sally will look in the box. They fail to realize that in the absence of new information, Sally will retain her (now false) belief that the marble is still in the basket – to use the common term, autistic children lack an adequate Theory of Mind.

In the first act of the puppet show, the child presumably believes the following:

(4) The marble is in the basket, and
(5) Sally BELIEVES that the marble is in the basket.

Then in act 2, the child learns that:

(6) The marble is in the box

and presumably updates her beliefs incorrectly to infer that:

(7) Sally BELIEVES that the marble is in the box

Recall the conclusions of the previous discussion: 1) second-order beliefs do not necessarily require metarepresentations (it is only necessary to have the ability to *represent* first order beliefs in order to *have* second-order beliefs), and 2) propositional embedding and referential opacity are required for second-order beliefs. Following from these conclusions, it seems clear that the Sally/Anne test does not imply an autistic deficit to do with second-order representations. Rather, it implies that either: 1) the autistic child does not have a concept of belief, or 2) the autistic child has a concept of belief but cannot handle the processing requirements of referential opacity and/or propositional embedding. In fact, the evidence quoted in (Leslie, 1991) is insufficient to distinguish between these two possibilities. Children are never directly asked about the beliefs of others ("Where does Sally *think* her marble is?") Rather, they are asked something like, "Where will Sally *look* for her marble?"

The second area of evidence quoted by Leslie is the apparent lack of pretend play in autistic children, and it is on this basis that he develops the metarepresentational conjecture and defines what he means by "metarepresentation". "I have used the term 'metarepresentation' in a specific sense: to mean (e.g., in the case of understanding pretence-in-others) an internal representation of an epistemic relation (PRETEND) between a person, a real situation and an imaginary situation (represented opaquely)…" (Leslie, 1991, p. 73) This definition doesn't sound at all like the definition of metarepresentation as higher-order representation pursued above. It seems like a highly technical redefinition of the word. This is a fact that Leslie seems to be quite aware of, as he says in a footnote that "'metarepresentation' can mean something like 'a kind of proprietary (internal) representation in ToM mechanisms' and something like 'a particular concept of representation which someone grasps'." (p. 77) It is not clear what Leslie's second possibility in the above refers to, but what he probably has in mind is Perner's (1991) account, which differs from both Leslie and Dennett. Unfortunately, he does not elaborate any further. From now on, I will call the definition of "metarepresentation" as higher-order representation "metarepresentation$_1$", while Leslie's version will be "metarepresentation$_2$" (and I'll forget about Perner's definition for the purposes of this discussion).

With that in mind, let's take a look at Leslie's formalism of the PRETEND example. In his view, the predicate PRETEND (which is supposed to behave similarly to BELIEVE and DESIRE) works something like this:

(8) Mother PRETEND the empty cup "it contains tea" (p. 73)

In addition to the new definition for metarepresentation$_2$, Leslie is also using a very different formalism for his psychological predicates – three arguments instead of two. Two questions immediately arise: 1) is Leslie's formalism plausible and/or compatible with the BELIEF/DESIRE formalism pursued above? and 2) putting aside Leslie's metarepresentation$_2$, is there anything metarepresentational$_1$ in his alternative formalism?

The Plausibility of Leslie's Formalism

Much of what I have to say in this section and the next parallels critiques from Pernerwith which I am in broad agreement (for example, Perner, 1991) . Rather than give a full analysis of Leslie's ideas, I will concentrate on the points I need to make for the discussion to follow.

The first observation is that there appears to be an important difference between pretending and believing, so we need to be cautious about generalizing from one to the other. Although it is possible to have beliefs without any representation of belief, it is not at all clear that this also holds for pretence. The possibility of pretending that something is true without being aware that one is doing so seems unlikely. So whereas to believe that the cup is empty does not require the self-conscious reflection that

(9) I BELIEVE that the cup is empty,

there is no way to pretend that the cup contains tea without self-conscious reflection by the subject on her own mental state. That is, the subject would have to BELIEVE:

(10) I PRETEND that (the cup contains tea).

Therefore, unlike beliefs and desires, being able to pretend seems to imply the ability to understand pretence in oneself, and thus in others, since the forms are the same. For instance believing that:

(11) Mother PRETENDS that (the cup contains tea)

requires exactly the same representational capacities as believing that one is pretending oneself.

Getting back to the substance of the issue, Leslie's formalism is actually quite different from the above. In his system, pretence is represented more like this:

(12) Mother PRETENDS (the empty cup) ("it contains tea")

That is, he has three elements: the subject (Mother), the real situation (the empty cup), and the pretend situation (it contains tea). But why is it that in order to understand pretence, you must be aware of exactly how the real situation differs from the imagined one? In reality, you can simply say "Mother pretends that the cup contains tea" and remain unsure of whether the cup is empty, contains orange juice, or whatever. That is, you do not need to know the "real" situation to understand the pretence. All you need to know is the fact, contained in the semantics of PRETEND with it's implied referential opacity, that the real situation must differ in some way from the imaginary one. If this is clear in the case of PRETEND, it is even more so in the case of BELIEVE. It would be much too restrictive to suppose that BELIEVE requires knowledge of the actual situation as in:

(13) Mélissa BELIEVES (her dog is dead) ("her dog is dead"), or
(14) Mélissa BELIEVES (her dog is **not** dead) ("her dog is dead")

Again, the information one needs is bound up in the semantics and referential opacity of BELIEVE. When you believe that *p*, *p* may or may not be true. Further problems arise when we try to embed Leslie's formalizations of psychological predicates to form second order beliefs. For instance, to believe (12) above would require:

(15) I BELIEVE [Mother may or may not PRETEND (the empty cup) ("it contains tea")]$_a$ ["Mother PRETENDS (the empty cup) ("it contains tea")"]$_o$

where the subscript "a" above marks the actual situation and "o" marks the referentially opaque proposition. This situation just seems unnecessarily complicated. You simply don't need to know the real situation in order to evaluate the truth of psychological predicates. The principle of referential opacity gives you everything you need to know – that the embedded proposition may or may not be true regardless of the truth-value of the psychological predicate.

Metarepresentation$_1$ in Leslie's Formalism

Is there anything metarepresentational$_1$ in Leslie's formulation of the semantics of psychological predicates? Leslie has made the unusual move of including the actual situation alongside the imagined situation in his formulation of at least one of the psychological predicates. Does this move change anything in the analysis of metarepresentations$_1$ in psychological predicates? The only real complication here is the introduction of dual representations for the same object – for example, the cup as an empty cup and the cup as a cup with tea in it. This dual representation

is well accounted for in Von Eckardt's (1999) definition of mental representation. The first refers to a concrete object and/or a property of a concrete object, while the second refers to an object/property in a possible or fictional world, or in the case of BELIEVE may simply represent an object/property incorrectly. So the dual representation does not imply metarepresentation$_1$.

One other aspect of Leslie's formulation deserves consideration. In the "Mother PRETENDS" example above, the first situation (the empty cup) is referred to again in the imaginary situation (it contains tea). The anaphoric reference in the imaginary situation ("it") could perhaps be taken to imply that the *representation* of the empty cup, rather than the empty cup itself, is the subject of the imaginary representation, thus making the latter a metarepresentation$_1$, but this is probably not the interpretation Leslie had in mind.[1] In fact it is hard to imagine how such an interpretation could be made coherent, since unpacking the imaginary situation would lead to "(the representation of the empty cup) contains tea" – and that can't be right.

Metaconfusion

Leslie is self-consciously using a technical definition for metarepresentation$_2$ that does not intersect in any way with Dennett's metarepresentation$_1$. Nevertheless, for other authors, the distinction may not be so clear. The potential for confusion is quite neatly demonstrated in a paper by Whiten and Byrne (1991). In an otherwise excellent article about the implications of Leslie's ideas for studies of pretend play in primates, they explicitly state that Leslie's metarepresentation$_2$ is second-order representation (i.e. metarepresentation$_1$). But the confusion doesn't stop there. They go on to offer a summary of Leslie's theory of metarepresentation$_2$ that is worth quoting at length.

"Leslie argues convincingly that the isomorphism between the properties of mental state terms and those of pretend play is not coincidental, but signifies a fundamental psychological achievement which can generate both pretence and an ability to represent the mental states of others. What these two share is that they are *representations of representations* – labeled variously as second-order representations (Dennett) or metarepresentations (Pylyshyn, Leslie).

"In the case of mental state terms, what 'second-order' means is fairly obvious: the child's mind represents a mental state in another's mind, *believing* (for example) that her father *thinks* there is a mouse behind the chair.

"In the case of pretence, the implication is less obvious. The key point is that in pretence, as strictly defined by Leslie, two simultaneous representations of the world must coexist in a precise relationship. When a child talks into a banana as if it were a telephone …the child has a primary representation of the object as a banana and, simultaneously, a representation of it as a telephone … The pretend representation is coded or marked off in some way as metarepresentational…" (Whiten and Byrne, 1991, p. 269, their italics.)

The first two paragraphs above demonstrate the confusion nicely. The authors are explicitly running together Dennett's metarepresentation$_1$ with Leslie's metarepresentation$_2$. Furthermore, they are committing the error of assuming that second order beliefs require second order representations. To see this, consider their example in the final paragraph, which makes use of the psychological predicate THINK. They make it seem like the child must have a representation of her father's thoughts, which of course consist of representations. Therefore the child must be engaging in second-order representation, or metarepresentation$_1$. But "thinks" in this context means the same thing as "believes", and so the appropriate formulation is actually:

(16) The child BELIEVES that her father BELIEVES that there is a mouse behind the chair.

This is straight-up second-order belief, which I have shown to *not* necessarily involve second-order representation, or metarepresentation$_1$.

The final paragraph picks up on the "real situation" vs. "imaginary situation" component of Leslie's formulation and reads into it another sense of "metarepresentation", which I'll call "metarepresentation$_3$" – definition: a representation of a counterfactual state of affairs. But counterfactual representations are fully compatible with the fairly non-controversial theory of first-order mental representations put forth by Von Eckardt (1999).

The confusion in Whiten and Byrne really comes to the fore in their concluding sections, where they talk about a "cluster of metarepresentational capacities." The first capacity they discuss is indirect sensorimotor coordination – the ability that humans and some other primates have to direct the actions of parts of their bodies by looking in a mirror or at a video image of the body parts they are trying to control. This, according to Whiten and Byrne, requires "a capacity to represent the remote representation of parts of self available in the mirror or video image: second-order representation" (p. 279). This ability is straightforwardly metarepresentational$_1$ in the Dennettian sense. In fact, it is a case of hybrid metarepresentation$_1$, requiring a mental representation of an external representation.

The other two "metarepresentational" abilities are tool use and insight. Tool use (in this case, a chimpanzee using a branch to probe for termites)

[1] Recall Leslie's definition above: "…an internal representation of an epistemic relation…"

apparently requires "a capacity to generate, simultaneously with the primary perception of the branch as branch, a metarepresentation of it as probe" (p. 280). Insight is a leap from pretence to re-description as in, "'I *pretend* this rock is a hammer' … 'Aha, I could *use* this rock as a hammer'...." (p. 280, Whiten and Byrne's italics). This is much closer to Leslie's technical definition of metarepresentation$_2$ in which the representation of the world as it really is coexists with a pretend representation of the world. But as I argued above, Whiten and Byrne appear to have drawn on the occurrence of a counterfactual in Leslie's formalism to build a third sense (metarepresentation$_3$), which is at work in the above.

In conflating metarepresentation$_1$ with their own interpretation of metarepresentation$_2$ (metarepresentation$_3$), Whiten and Byrne have made two mistakes, one of which comes directly from Leslie, and one that is not explicitly present in (Leslie 1991). In the former case, they have imported Leslie's notion that psychological predicates require an explicit representation of how the world actually is in addition to the representation of how the world is believed, pretended, or desired to be, and used it to unwittingly arrive at a new definition of metarepresentation$_3$. But they have also made another mistake in equating Leslie's metarepresentation$_2$ with Dennett's metarepresentation$_1$, even to the point of citing Dennett and Leslie in the same sentence.

To be fair to Whiten and Byrne, their dissection of Leslie is itself an attempt to criticize and make some new distinctions. For instance, they point out that not all pretend play involves a real object. Humans and other apes appear quite capable of having imaginary friends, and interacting with imaginary objects. In this case, it is difficult to see what the "real situation" component of Leslie's formulation would amount to, and is evidence for at least sometimes abandoning it in favor of two-place intentional predicates. But confusion over the two original senses of metarepresentation, and the unwitting introduction of yet a third sense really manages to confuse the issue. For instance, in summing up, they speculate that perhaps, "what convinces those who interact intensively with them that chimpanzees are 'intelligent' is a facility in second-order representation." (p. 280) This is a nice parsimonious account, but it is built by equating three different definitions of metarepresentation based on a number of confusions about the nature of psychological predicates. As I have attempted to show, second-order beliefs and desires as well as the pretend play studied in Whiten and Byrne's work require propositional embedding and referential opacity, but do not necessarily require second-order representations (metarepresentation$_1$).

Conclusions and Prospects

In Dennett's *Making Tools for Thinking* (Dennett, 1998), he invites us to speculate along with him on the difference between what he terms "florid" and "pastel" representations. Florid representations are those that become explicit as objects in the world, by being encoded in language or some other physical medium (drawings on paper, for instance.) He notes that the capacity to form florid representations seems to imply the ability to manipulate the representations themselves, which leads him to raise the slogan "no florid representation without metarepresentation." He further speculates that "belief about belief" may not be the same thing at all as "thinking about thinking" – that is, having the ability to self-consciously reflect, compare notes with other thinkers, and so on. The considerations in this paper may help to shed a little light on all of these questions.

If I am right that second-order belief does not require metarepresentation$_1$, and Dennett is right that thinking about thinking requires florid representations and therefore metarepresentations$_1$, then maybe we do have the basis for a nice account of one possible difference between humans and other apes – a capacity to form and manipulate higher-order representations (that is, metarepresentations$_1$).

References

Baron-Cohen, Simon (1991). Precursors to a theory of mind. Andrew Whiten (Ed.), *Natural Theories of Mind: Evolution, Development, and Simulation of Everyday Mindreading*. (pp. 233-252). Oxford: Blackwell.

Dennett, Daniel (1998). Making tools for thinking. Dan Sperber (Ed.) (2000). *Metarepresentation*. New York: Oxford University Press.

Leslie, Alan (1991). The theory of mind impairment in autism. Andrew Whiten (Ed.) *op cit. (*pp. 63-78)

Perner, Josef (1991). *Understanding the Representational Mind*. Cambridge, Massachusetts: MIT Press.

Von Eckardt, Barbara (1999). Mental representation. Robert A. Wilson and Frank C. Keil. *The MIT Encyclopedia of the Cognitive Sciences*. (pp. 527-529). Cambridge, Massachusetts: MIT Press.

Whiten, Andrew, and Byrne, Richard W. (1991). The emergence of metarepresentation in human ontogeny and primate phylogeny. Andrew Whiten (Ed.) *op cit.* (pp. 267-282.)

Connectionist modelling of surface dyslexia based on foveal splitting: Impaired pronunciation after only two half *pint*s

Richard Shillcock[*†]
Padraic Monaghan[*]
rcs/pmon@cogsci.ed.ac.uk
[*] Institute for Adaptive and Neural Computation, Division of Informatics
[†] Department of Psychology
University of Edinburgh, 2 Buccleuch Place, Edinburgh, EH8 9LW, UK

Abstract

In cases of surface dyslexia and phonological dyslexia there is a dissociation between the reading of irregular words and nonwords. This dissociation has been captured in connectionist models of dyslexia in terms of impairments to the models' phonological representations. We report a series of connectionist simulations based on an alternative neuro-anatomically motivated theory that dyslexia is at least partly caused by hemispheric desynchronisation. Problems of interhemispheric transfer affect the mapping between orthography and phonology because the human fovea is precisely vertically split: a fixated word is initially split and the two parts contralaterally projected to the two hemispheres of the brain. Much of lexical processing can be reconstrued as the integration of this information (Shillcock, Ellison & Monaghan, 2000). We demonstrate that the dissociation between the reading of irregular words and nonwords can be understood in terms of a failure to integrate the initially split input.

Introduction

The word *pint* is pronounced differently from its orthographic neighbours *dint*, *hint*, *lint*, *mint*, *tint*. In this sense its pronunciation is irregular. Because of the presence of such words in its lexicon, English is said to have a deep orthography. Irregular pronunciations have been observed to cause processing problems both for normal readers and for certain impaired readers: in the former a low frequency irregular word such as *pint* may take measurably longer to pronounce than a matched regular word, and in the latter such a word might be erroneously regularised (*pint* pronounced to rhyme with *mint*). For these reasons irregular words have been used in a large number of studies of how readers translate the orthographic representation of a word into its phonological specification. In this paper we approach the pronunciation of irregular words from the perspective of a new model of visual word recognition based on the neuro-anatomical observation of foveal splitting (Shillcock, Ellison & Monaghan, 2000). We will argue that this model provides a principled motivation for some of the problems associated with irregular pronunciations in both normal and impaired readers.

The measurable processing problems that normal adult readers have with low frequency irregular words (*pint*) do not extend to high frequency irregular words such as *have* (whose pronunciation cannot be predicted from *cave*, *rave*, *pave*, *knave*) (Paap & Noel, 1991; Seidenberg, Waters, Barnes & Tanenhaus, 1984; Taraban & McClelland, 1987). Perhaps the most productive approach to modelling this interaction between regularity and frequency has been that developed by Seidenberg and McClelland in their (1989) connectionist model of reading and by Plaut and others subsequently (e.g. Plaut & McClelland, 1993; Plaut, McClelland, Seidenberg & Patterson, 1996; Harm & Seidenberg, 1999). Seidenberg and McClelland's original model was a simple feedforward model; later models have contained recurrent connections and more structured input and output layers. In general, in these models the apparently rule-based behaviour that generates regular pronunciations is recast in statistical terms, so that a low frequency irregular pronunciation (*pint*) is militated against by the more frequently occurring regular pronunciation of the relevant vowel in similar contexts. Thus, when the irregular pronunciation is itself high frequency (*have*), that particular mapping between orthography and phonology is sufficiently emphasised to co-exist with the regular mapping. The superpositional storage that characterises these models means that a minority mapping such as that required by *pint* is disproportionately demanding in terms of computational resources compared with the efficient generalisation represented by a regular pronunciation. This fact is also demonstrated in Plaut and McClelland's (1993) model in which attractors behave componentially when the input and output representations of words are split into onset, nucleus and coda. In their model, pronunciation of the vowel in *pint* involves the connections from all three input slots (onset, nucleus and coda), whereas the regular, default pronunciation of the *i* involves only the nucleus.

Irregular words pose particular problems for surface dyslexics, who are liable to produce regularisation errors (Patterson, Coltheart & Marshall, 1985; Manis et al., 1996). In connectionist models, the learning and retention of irregular pronunciations are generally vulnerable. For instance, Seidenberg and McClelland showed that restricting the number of hidden units impairs learning low frequency irregular words, and Harm and Seidenberg (1999) produced a similar effect by lowering the learning rate overall. Plaut et al. (1996) and others explore the idea that a division of labour occurs between the direct orthography-phonology mapping and the same

mapping mediated by semantics: the irregular pronunciations are seen as relying more on the route that proceeds via semantics, leaving the direct route to concentrate on the regular pronunciations. This behaviour of the models seems to capture the observation that developmental dyslexics frequently have impaired phonological processing (see Snowling, 2000, for a comprehensive review of the phonological processing problems of dyslexics).

Some of the most critical data in this area concerns the dissociation between the ability to pronounce irregular words and the ability to pronounce nonwords. This dissociation is found between the surface and phonological subtypes of dyslexia (Beauvois & Derouesné, 1979). Surface dyslexics cope moderately well with novel words and nonwords, but are liable to make regularisation errors on known irregular words, whereas phonological dyslexics cope moderately well with irregular words but are disproportionately impaired when reading novel words and nonwords. This dissociation motivated Marshall and Newcombe's (1973) original dual-route model and its later development and computational implementation by Coltheart and others (e.g. Coltheart, Curtis, Atkins & Haller, 1993), in which a lexical route and a non-lexical route (containing grapheme-to-phoneme correspondence rules) can be differentially impaired to produce the desired impairments.

In this paper we claim that developmental surface dyslexia, characterised by problems with irregular words, arises naturally from impaired hemispheric interaction in a model based on the observation that the human fovea is precisely vertically split. There are long-standing observations that dyslexia is frequently associated with problems of callosal transfer (e.g. Davidson & Saron, 1992). By modelling word recognition within a split network, we ground these observations of impaired reading in an implemented model of normal reading. We claim that impaired hemispheric interaction is a fundamental, qualitative explanation of problems in pronouncing irregular words, and is a more parsimonious account than resource-based explanations.

The split-fovea model of reading

Shillcock, Ellison and Monaghan (2000) present a model of lexical access based on the precise vertical splitting of the human fovea. Information presented in the left visual field (LVF) projects, initially, to the right hemisphere (RH), whereas information in the RVF projects to the LH. This long-recognised initial contralateral projection of the visual field to the two hemispheres of the brain is also true of the human fovea – a fact that has not been extensively explored in research in visual word recognition. When a word is directly fixated, the two parts of the word on either side of the fixation point are projected to different hemispheres. In order for a word to be recognised and pronounced correctly, the information in the two hemispheres has to be integrated. Shillcock et al. (2000) investigate some of the implications of foveal splitting for a full-sized lexicon and show that the initial splitting of the word is an informationally attractive starting point for word recognition, rather than being merely an inconvenience.

Consider the word *pint*, centrally fixated. The two sides of a split model will receive the two letters *pi* and *nt*, respectively. In order to pronounce the vowel correctly, the model must process information from both sides of the model: *pi* on its own may be pronounced as in *pine*, or as in *pill*. If this integration is not complete then a regularisation error is likely to occur, so that *pi* will be pronounced in its most frequent form: /pI/ (see Table 1). As Harm and Seidenberg (1999) observe, the task of reading irregular words is akin to solving the XOR problem. In the case of a split model without recurrence, this task is impossible, as the structure is akin to a perceptron.

Table 1: Pronunciation of *pi*.

Pronunciation of vowel	Example	Count	Frequency (per million)
/I/	pith	14	424
/ii/	piece	2	175
/&I/	pint	10	164
/I@/	pier	2	16

When a single word is read it may be fixated to the left of the first letter, to the right of the last letter, or at all points in between. Elsewhere (e.g. Shillcock & Monaghan, *2001*) we have implemented an idealised version of the initial splitting of these different visual inputs in reading single words in a series of neural network models, so that there are five different fixation positions across the input for a four letter word (only "fixations" between letters are considered). In the simulations we report here, we employ a simplified version of this model, in which each word is only fixated at one fixation point within the word. This simplification allowed us to stay closer to Harm and Seidenberg's (1999) simulations, which provide an important point of comparison. We show that impairments to the integration of information in the two parts of the word results in the behaviour associated with surface dyslexia.

One version of the split model of reading is shown in Figure 1. The model comprises two orthographic input layers, corresponding to the left and right visual fields. Each input layer has 4 letter slots. If a letter is present in a slot then one of 26 units representing the letters of the alphabet will be active. The output layer is a representation of phonology, with two slots each for onset, nucleus and coda. Phonemes are represented in terms of 11 phonological features. We have used the features described by Harm and Seidenberg (1999), although we have augmented their phonology to accommodate the transcriptions found in the CELEX database (Baayen, Pipenbrock & Gulikers, 1996). These changes to the phonological transcription principally involved the representation of diphthongs and the role of schwa; the changes consid-

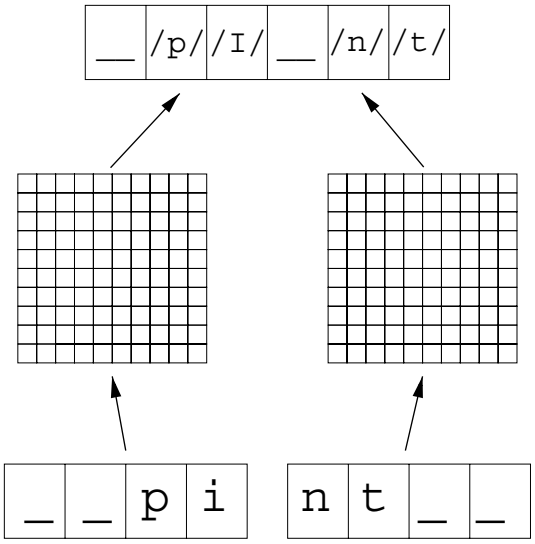


Figure 1: The split model of reading.

erably increased the problems of learning the mapping from orthography to phonology. Each input layer in the model is fully connected to one of two hidden layers each with 100 units. These hidden layers are fully connected to the output. The model has to learn to map orthography onto phonology given the constraints of a split input.

The model was trained on 3835 single-syllable word lemmas from the CELEX English database with up to two consonants in the onset and coda. Four words were omitted because they contained orthographic consonant clusters that would not fit within the input[1]. Words were presented in the input so that the first vowel was justified immediately to the left of centre. In total, 57.5% of words were fixated at or to the left of their physical centre, which is the usual site of fixation during reading (Rayner, 1998). Words were presented randomly to the model according to the log-frequency of the word, and backpropagation was used to train the connections between the layers.

Several versions of the model were trained: (a) a feedforward version with no phonological attractors ("Split-No Attractors"), (b) a recurrent version with phonological attractor units ("Split-Attractor"), and (c) a recurrent and interconnected version with phonological attractors, with the hidden layers fully connected to one another ("Split-Interconnected"). The attractor models had a set of 35 "clean-up" units connected to the output layer, which were pretrained in learning a phonology to phonology mapping. Orthography to phonology trials were then interdispersed with phonology to phonology trials during training of the attractor models. Two nonsplit models were also trained as controls, equivalent to those employed by Harm and Seidenberg (1999). These models were required to make the same mapping between orthography and phonology, except that all the input units were connected to a single hidden layer of 100 units. Feedforward ("Nonsplit-No Attractor") and recurrent attractor ("Nonsplit-Attractor") versions were trained.

The 361 nonwords used by Harm and Seidenberg (1999) were employed[2]. The model was tested with 44 irregular words taken from Taraban and McClelland's (1987) materials[3].

We predicted that the Nonsplit-Attractor model would perform well on both irregular words and nonwords, following Harm and Seidenberg's demonstration of the capabilities of that model. The Nonsplit-No Attractor version should perform relatively poorly on irregular words and nonwords; this model corresponded to Harm and Seidenberg's (1999) unimpaired model, and their phonologically impaired model, respectively). We predict that the split models will exhibit surface dyslexia to varying degrees according to the level of interaction between the two hemispheres. Furthermore, this deficit will be robust in the face of further training on the model – additional training will not reverse the pattern of difficulties in reading irregular words and nonwords, as happens in Harm and Seidenberg's (1999) delayed model of reading.

Results

Figures 2, 3 and 4 show the performance of the different models, in terms of percentage of words correctly pronounced, at different stages of training. Figure 2 shows how well the models learned the whole training set, and we see that the NonSplit-No-Attractors model performs comparably to the same architecture presented by Harm and Seidenberg, climbing steadily to levels in excess of 90% correct. Even though the current training set contained more elaborate phonological representations than those used by Harm and Seidenberg for the same type of model, we see that the curve has not asymptoted even after the presentation of 5M words in training; further training promises to improve performance even more. We see a similar level of success for the NonSplit-Attractors model, though this model is slower to learn than the NonSplit-No-Attractors model due to the interleaving of phonology to phonology trials during training. These models replicate Harm and Seidenberg's (1999) success with the same architectures.

In contrast, Figure 2 shows that the Split-No-Attractors model asymptotes early, after about 3M words of training, and never exceeds 70% of words correct. The simple split model is fundamentally incapable of more than this modest performance. The remaining learning curves in Figure 2 demonstrate the value of connectivity between the two halves of the split model: the Split-Attractors model behaves on a par with the different

[1]The omitted words were *eighth*, *borscht*, *touched*, and *schnapps*.

[2]Three nonwords were omitted because they appeared as real words in our training corpus: *plop*, *mo*, and *peep*.

[3]4 of the original 48 items were omitted because they were wordforms that did not occur in our word lemma training set: *does*, *said*, *says*, and *were*.

Non-Split models, and the most rapid learning of all occurs in the Split-Corpus-Callosum model, although the latter model is not directly comparable with the others as it contains more weighted connections and hence more computational resources. The principal result from the training of the different models is that sharing information between the two halves of the input is critical to successful learning.

Figure 3 shows relatively successful generalisation by all of the models to the set of nonwords. The Split-No-Attractors model performs least well as we might predict from Figure 1, but all of the models asymptote within the 55%-70% region, in generalising to pronounce nonwords that were not encountered in training.

The central result of all of these simulations can be seen by comparing Figures 3 and 4. In Figure 4 we see differences between the models in their performance on pronouncing irregular words. The Split-No-Attractors model performs extremely poorly on these words, pronouncing only about 50% of the irregular words correctly. The Split-Attractors and NonSplit-Attractors models perform comparably well, and the NonSplit-No-Attractors and Split-Corpus-Callosum models perform extremely well on these irregular words. In relation to the performance of the other architectures, the simple split architecture shows a dramatic dissociation between

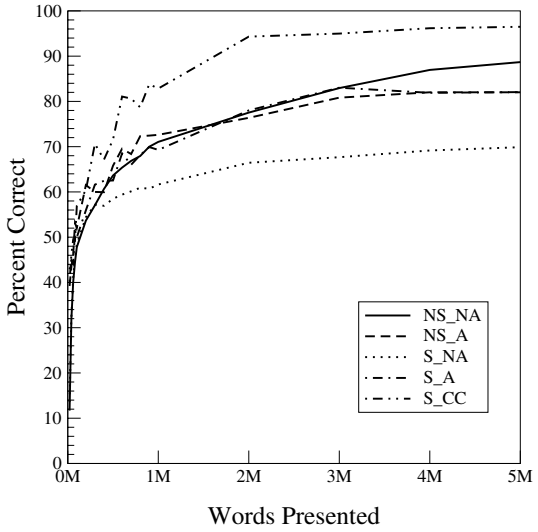


Figure 2: Performance of the split and nonsplit models on the training set.

The errors produced by the simple split model resembles those found in surface dyslexia. Table 2 compares the performance of the Split-No Attractors and NonSplit-No-Attractors models on an example set of irregular words after 5 million words had been presented to the models. The NonSplit model converges to correct pronunciations for all these words, whereas the Split model

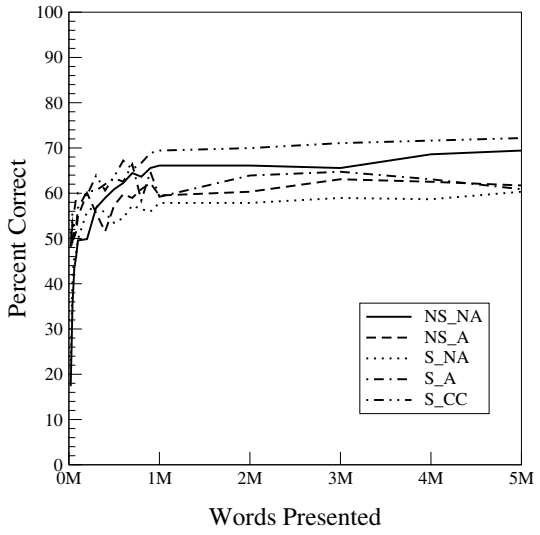


Figure 3: Performance of the split and nonsplit models on nonwords.

produces regularisation errors even after such lengthy training. The differences between the Split and NonSplit models are not only quantitative but qualitative, with the Split model producing plausible over-regularisations for irregular words.

Table 2: Productions of the Split and NonSplit No-Attractor models for irregular word examples.

	Pronunciation		
Word	Correct	NonSplit model	Split model
bind	b&Ind	b&Ind	b&nd
broad	brOOd	brOOd	brOOd
come	kVm	kVm	kVm
hood	hUd	hUd	huud
mild	m&Ild	m&Ild	mEIld
pear	pE@r	pE@r	pIEr
pint	p&Int	p&Int	pInt
quay	kii	kii	kEI
tomb	tuum	tuum	tOm

Discussion

We have successfully reproduced the dissociation between the reading of irregular words and nonwords observed in surface dyslexia. We started by observing that the human fovea is precisely vertically split, and that a fixated word is initially divided between the two hemispheres. We reconstrued the task of lexical processing as one of integrating the information contained in the two hemispheres. We explored the performance of neural network models of reading which had been similarly vertically split, and demonstrated that simple split archi-

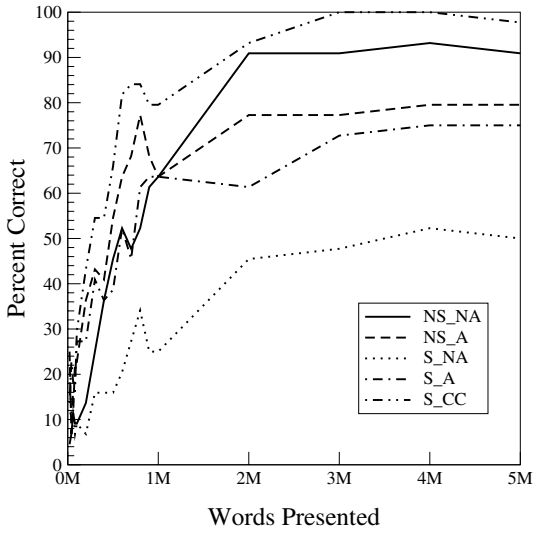


Figure 4: Performance of the split and nonsplit models on irregular words.

tectures are crucially limited in their success in mapping from orthography to phonology.

The problem of pronouncing a divided irregular word like *pint* resembles the XOR problem. The two halves cannot map directly and independently onto the output to solve the problem. There is no purely componential solution, as there is with the pronunciation of a regular word like *mint*. Instead, there have to be intermediate representations that combine the two halves and then map onto the eventual solution. The problem is not restricted to the pronunciation of vowels, although they are the main part of the problem. The consonant cluster at the beginning of *chef* is liable to be regularised to resemble that in *chad* and *cheat* if the whole of the word is not present.

A paradox of the connectionist modelling of cognitive neuropsychological data is the emergence of complex dissociations from relatively undifferentiated architectures. Such demonstrations are parsimonious accounts of the data because they seem to emerge from the structure of the problem, "out there" in the world rather than from the details of some putative functional processing architecture. Researchers have produced successful connectionist accounts of the pattern of dissociations found in the different dyslexias. These dissociations have been achieved in a variety of ways, and we can distinguish three broad strategies. The first is the *Representational Strategy*. Thus, in Plaut and Shallice's (1993) model of deep dyslexia the authors show that the proportion of the different types of error produced by the lesioned network can be changed by lesioning in different parts of the model; lesioning around the model's semantic clean-up units causes more semantic errors, for instance. Similarly, Harm and Seidenberg (1999) show that impairing the phonological attractors in their model increased the errors on irregular words. The psychological reality of this strategy is relatively easy to assess. For instance, Harm and Seidenberg's impairing of the phonological attractors is motivated by the data showing that surface dyslexia is often accompanied by phonological impairment (see, e.g., Snowling, 2000). The second strategy we identify is the *Parametric Strategy*. Examples are the manipulation of the computational resources available to the model, or changes in the details of training. Seidenberg and McClelland (1989) reduced the numbers of hidden units available to the orthography-to-phonology mapping, showing that it militated against learning the irregular words. Similarly Harm and Seidenberg (1999) produced the same outcome by reducing the learning rate. Such manipulations are quantitative, compared to the qualitative effects of the Representational Strategy, and their psychological reality is more difficult to assess. There are clear demonstrations that the orthography-to-phonology mapping for irregular words is the hardest aspect of the pronunciation problem, but equally there are demonstrations that the problem can be solved by a network model with only a few hundred nodes, given the correct representations. Are irregular words still hard for a processor with the resources of the human brain, and with several years to spend on the problem? Additional training of Harm and Seidenberg's model with a low learning rate producing surface dyslexia behaviour leads to the convergence of performance on regular and irregular words, for example. Parametrically based models of dissociations in dyslexia carry with them costly assumptions concerning the capacities of the human brain.

Our own approach has been to introduce a discrete, qualitative neuro-anatomical distinction into the modelling: the effects of foveal splitting. We characterise surface dyslexia as being caused, at least in part, by hemispheric desynchronisation. No amount of extra training and no amount of extra computational resources devoted to either half of a split processor can improve performance on irregular words in a direct mapping between orthography and phonology. The problem is a qualitative one.

Thus, we have provided an account of the dissociation between irregular words and nonwords in surface dyslexics. We do not see this hemispheric desynchronisation account as a necessarily exclusive one. There may also be a contribution from the impairment of phonological representations, as is generally assumed. The relative contribution of each account, and any possible interaction between them, is an empirical question. However, it may be possible to ground phonological impairment itself in hemispheric desynchronisation, despite the fact that expressive phonology is conventionally viewed as the sole preserve of the LH.

A further aspect of the dissociation between regular words and nonwords concerns the performance of phonological dyslexics, who can be highly proficient readers of known words, but can also be dramatically poor at reading nonwords and unknown words, even to the extent of not being able to generate the sound of an isolated letter. We interpret phonological dyslexia as also

resulting from hemispheric desynchronisation. We propose that the desynchronisation is more severe in phonological dyslexia, compared with surface dyslexia, and that a different relationship emerges between the orthographic, phonological and semantic forms of words to compensate for the inability to integrate the orthographic information in the two hemispheres and to map it directly onto a phonological representation. We propose that orthographic information is mapped directly onto semantic information independently in the two hemispheres; each hemisphere partially activates the semantic representations of all the words corresponding to its own orthographic input. Identification of the word is achieved by the interhemispheric transfer of semantic information. The routes by which semantic information might be transferred interhemispherically are more extensive, compared with those concerned with vision, and the problem of finding the intersection of two sets of partially activated semantic representations is, we claim, an easier problem than integrating the corresponding visual information. In this account, a novel word can only be pronounced ad hoc by analogy with known words. This account is therefore a qualitative explanation of the often dramatic problems with novel words observed in phonological dyslexia.

Acknowledgments

This work was supported by Wellcome Trust grant 059080.

References

Baayen, R.H., Pipenbrock, R. & Gulikers, L. (1995). *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Beauvois, M.-F. & Derouesné, J. (1979). Phonological alexia: Three dissociations. *Journal of Neurology, Neurosurgery and Psychiatry*, *42*, 1115–1124.

Coltheart, M., Curtis, B., Atkins, P. & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100*, 589–608.

Davidson, R.J & Saron, C.D. (1992). Evoked potential measures of interhemispheric transfer time in reading disabled and normal boys. *Developmental Neuropsychology 8*, 261-277.

Harm, M. W. & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: insights from connectionist modelling. *Psychological Review*, *106*, 491–528.

Manis, F., Seidenberg, M., Doi, L., McBride-Chang, C. & Peterson, A. (1996). On the basis of two subtypes of developmental dyslexia. *Cognition*, *58*, 157–195.

Marshall, J.C. & Newcombe, F. (1973). Patterns of paralexia: A psycholinguistic approach. *Journal of Psycholinguistic Research*, *2*, 175–199.

Paap, K. R. & Noel, R. W. (1991). Dual-route models of print to sound: Still a good horse race. *Psychological Review*, *53*, 13–24.

Patterson, K., Coltheart, M. & Marshall, J.C. (1985). General introduction. In J.C. Marshall, M. Coltheart & K.E. Patterson (Eds.) *Surface dyslexia*. Hillsdale, NJ: Erlbaum.

Plaut, D.C. & McClelland, J.L. (1993). Generalization with componential attractors: word and nonword reading in an attractor network. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Erlbaum.

Plaut, D.C., McClelland, J.L., Seidenberg, M.S. & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*, 56–115.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*, 377–500.

Rayner, K. (1998) Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 373–422.

Seidenberg, M. S. & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523–568.

Seidenberg, M. S., Waters, G. S., Barnes, M. A. & Tanenhaus, M. K. (1984). When does irregular spelling or pronunciation influence word recognition? *Journal of Verbal Learning and Verbal Behavior*, *23*, 383–404.

Shillcock, R. Ellison, M. T. & Monaghan, P. (2000). Eye-fixation behaviour, lexical storage and visual word recognition in a split processing model. *Psychological Review*, *107*, 824–851.

Shillcock, R. & Monaghan, P. (1998). Using physiological information to enrich the connectionist modelling of normal and impaired visual word recognition. *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 945–950). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shillcock, R. & Monaghan, P. (2001). The computational exploration of visual word recognition in a split model. *Neural Computation*, *13*.

Snowling, M. (2000). *Dyslexia: A Cognitive Developmental Perspective* (Second Edition). Oxford: Blackwell.

Taraban, R. & McClelland, J. L. (1987). Conspiracy effects in word pronunciation. *Journal of Memory and Language*, *26*, 608–631.

Assessing Generalization in Connectionist and Rule-based Models Under the Learning Constraint

Thomas R. Shultz (shultz@psych.mcgill.ca)
Department of Psychology; McGill University
Montreal, QC H3C 1B1 Canada

Abstract

Although it is commonly assumed that rule-based models generalize more effectively than do connectionist models, the comparison is often confounded by pitting hand-written rules against learned connections. Three case studies from cognitive development show that, under the constraint that both types of models learn their representations from equivalent examples, generalization is consistently superior in connectionist models.

Generalization Problems

A significant part of the ongoing debate between rule-based and connectionist modeling in psychology has focused on the ability to generalize. A common claim from supporters of the classical, symbolic approach is that rule-based models are superior because they generalize more effectively than do connectionist models (Ling & Marinov, 1993; Pinker, 1997; Marcus, 1998). Generalization is considered important by most modelers because it distinguishes understanding of a problem from mere memorization of solutions.

The generalization ability of rules is often enhanced by the use of variables that can be bound to any number of objects or events. Consider the following rule, written in Common Lisp for a production system program. It generates correct responses on some Piagetian conservation of number problems:

```
((response more ?x ?y)
 (and
  (initially-same-number ?x ?y)
  (or (add1 ?x)
      (subtract1 ?y))))
```

The rule says to conclude that row *x* has more items than row *y* if the two rows initially had the same number of items and if one item was subsequently either added to row *x* or subtracted from row *y*. It has plenty of generality because the variables *x* and *y* can be bound to any rows with any number of items. It could be made even more general by adding a third variable *n*, representing the number of items added or subtracted.

More generally, a rule can be defined as a conditional statement in which conjunctively and disjunctively connected conditions, if verified as true, produce a set of conjunctively connected conclusions. Each condition and conclusion is a proposition that can be stated in predicate-argument form, where arguments can be constants, variables, or other propositions.

Leaving aside the issue of whether people actually generalize as well as such rules do, the claim has commonly been made that connectionist models rarely learn to generalize that well. Indeed, this argument seems to have been accepted by many connectionists (e.g., Anderson, 1995), and is at least partly responsible for the many attempts to improve generalization in neural network learning (e.g., Reed & Marks, 1995).

However, closer inspection reveals a serious confound in this argument. The symbolic rules are often written by hand, or perhaps merely alluded to, while the neural network learns its own connection weights by processing examples. The purpose of this paper is to remove this confound between representation and learning by requiring both types of model to learn their representations from equivalent examples. It is already well known that an alternate method of removing this confound, by hand-designing neural networks to explicitly implement rules and variables, also produces excellent generalization (Shastri, 1995).

The learning constraint proposed here is reminiscent of the developmental tractability constraint proposed by Klahr (1984). In discussing cognitive development, Klahr argued that any two plausible, consecutive developmental states must be integrated in a transition theory that can transform one state into the other. Similarly, and a bit more generally, I propose a constraint that acquired knowledge representations, whether rules or weight vectors, must be learned by a model in order to be considered plausible. Knowledge representations that are instead hypothesized to be produced through biological evolution may be dealt with by hand designing rule-based and connectionist models as noted earlier, or even more ambitiously, by simulated evolution. Covering both inherited and acquired representations is the more general principle that other, non-representational features must be held constant when assessing generalization ability. Otherwise claims about superior generalization ability may be confounded with acquisition issues and possibly other differences.

Choice of Algorithms and Domains

A systematic test of generalization under a learning constraint should eventually involve many algorithms and problem domains. To begin this process, this paper compares one leading connectionist algorithm to one leading rule-learning algorithm in three different domains of cognitive development.

One of the most frequently used connectionist algorithms in cognitive development, and the principal one used in my laboratory, is cascade-correlation (CC). CC creates feed-forward networks by recruiting new hidden units that correlate well with network error and installing them in cascaded layers (Fahlman & Lebiere, 1990). It has been used to simulate a wide variety of cognitive developmental phenomena, including conservation (Shultz, 1998), seriation (Mareschal & Shultz, 1999), the balance scale (Shultz, Mareschal, & Schmidt, 1994), shift learning (Sirois & Shultz, 1998), pronoun acquisition (Oshima-Takane, Takane, & Shultz, 1999), infant familiarization to rule-governed sentences (Shultz & Bale, 2001), and integration of the concepts of velocity, time, and distance for moving objects (Buckingham & Shultz, 2000).

Choosing an equivalent rule-learning algorithm encounters the problem that there are not all that many successful rule-based models of cognitive development, in the sense of implementing developmental transitions. A good case can be made that the largest number of successful rule-based developmental models have been achieved by the C4.5 algorithm (Quinlan, 1993) and its immediate predecessor ID3 (Quinlan, 1986). These include models of English past tense morphology (Ling, 1994; Ling & Marinov, 1993), the balance scale (Schmidt & Ling, 1996), grammar learning (Ling & Marinov, 1994), and reading (Ling & Wang, 1996). There is also a simulation of non-conscious acquisition of rules for visual scanning of a matrix (Ling & Marinov, 1994), and numerous applications in engineering and decision support (Quinlan, 1993). Among alternative symbolic rule-learning algorithms applied to the same phenomenon, the balance scale, C4.5 produced an arguably superior model.

C4.5 learns to classify examples described with features and values by forming a smallish decision tree that can be converted into production rules. It is a greedy (i.e., non-backtracking) algorithm that repeatedly finds the most informative feature with which to classify so far unclassified examples.

There are a number of intriguing similarities between C4.5 and CC. Both algorithms use supervised learning of examples, focus on largest current source of error, gradually construct a solution based on what is already known, and aim for a small solution that generalizes well. In this paper, I report on generalization performance of the CC and C4.5 algorithms on the three problems of conservation acquisition, number comparison, and infant familiarization to sentences in an artificial language.

Conservation Acquisition

A recent CC model of conservation acquisition focused on Piaget's conservation of number problems (Shultz, 1998). In one version of these problems, a child first agrees that two rows have the same number of items, and is then asked which row has more after one of the rows is transformed, for example, by compression. Children below about six years of age typically judge the longer row to have more items, whereas older children correctly judge the rows to remain equal. The vast psychological literature on conservation (over 1000 studies) has produced a number of well-replicated regularities. Among them are acquisition (with a sudden jump in performance), the problem size effect (with better performance and earlier success on small number problems than on large number problems), length bias in pre-conservation children (choosing the longer row as having more), and the screening effect (with young children giving a correct answer to a screened transformation until the screen is removed).

CC networks were trained on 420 examples of number conservation problems of row lengths and densities ranging between 2 and 6, with number of items being the product of length and density. Using inputs coding the length and density of each row, both before and after the transformation, the identity of the transformed row, and the identity of the transformation (addition, subtraction, compression, and elongation), networks learned to judge whether the rows had equal numbers or not, after the transformation. Both equal and unequal initial rows were included. Length and density were coded as real numbers, and the other inputs were coded in a localist binary fashion. There were 100 test problems of the same type, not used in training, to assess generalization performance.

C4.5 was trained with the same examples, learning to classify them into three numerical judgments: one row has more, the other row has more, or both rows have the same. C4.5 was equipped with ability to deal with continuous, as well as qualitative inputs,[1] and to use the option for information gain ratio, which is generally superior to simple information gain (Quinlan, 1993).

Proportions correct on training and test problems, respectively, were 1.0 and .95 for 20 CC networks, and .40 and .35 for 20 C4.5 trees. For both algorithms, generalization performance (on the test problems) was just a bit worse than performance on the training problems; but training and generalization performance was much higher for CC than for C4.5. If the learned

[1] C4.5 finds the gain ratios for each possible cutoff on a continuous feature and then chooses the partition of examples with the highest gain ratio in the usual way.

knowledge representation is inadequate, it does not afford good generalization. This makes a pure test of generalization ability difficult. To control for learning success, proportion correct on the test problems can be divided by proportion correct on the training problems, creating a generalization ratio. This ratio is .95 for CC and .87 for C4.5.

Because a failed model is not by itself very meaningful, I adopted the strategy of changing the input coding to C4.5 until learning was successful and then evaluating what is required to learn in terms of both theoretical plausibility and psychological coverage.

Following the lead of other C4.5 modelers (Schmidt & Ling, 1996), I coded the length and density input in relational, rather than absolute terms. For example, was the first row longer or shorter or the same length as the second row? Although this relational coding produced 100% success on training and test problems, it created knowledge representations that are unlike any that have been reported with children. For example, an English gloss of one of the smaller rules is: *If the first row is longer than the second row before the transformation, and shorter than the second row after the transformation, then the first row has more items.*

Because of this exclusive focus on relative length and density of the rows, there was never any reference to information on the transformation or the identity of transformed row. Nor could the C4.5 models cover any of the various psychological regularities. This is in distinction to both the CC model and children, characterized by a shift from concern with how the rows look to the nature and identity of the transformation. The CC model also covers all of the psychological regularities mentioned: sudden jump in acquisition, problem size effect, length bias, and the screening effect. Thus, although relational input coding can produce perfect learning and generalization in C4.5, it creates implausible knowledge representations and fails to cover the psychological data. In contrast, the CC model can learn and generalize effectively from raw input coding, acquire knowledge representations that are similar to those seen in children, and cover the psychological regularities.

Number Comparison

One of the most basic of numerical skills is that of comparing the size of two numbers. Prominent psychological regularities in number comparison are the min and distance effects. The min effect refers to earlier success and quicker performance the smaller the smaller of the two numbers. The distance effect refers to earlier success and quicker performance the larger the absolute difference between the two numbers.

My simulations focus on pairs created from the integers 0-9. In a study of interpolation, a randomly selected 50 pairs comprised the training set and the remaining 50 pairs comprised the test set. The integers were coded as real numbers, and there were three discrete output classes, including ties. Mean proportion correct on training and test problems, respectively, over 20 runs was 1.0 and 1.0 for CC and .75 and .66 for C4.5. The mean generalization ratio of test correct to train correct was 1.0 for CC and .89 for C4.5. Not only did CC learn the problem and generalize more effectively than did C4.5, but only CC captured the min and distance effects.

Knowledge representation analysis revealed a sensible solution for CC networks that involved positioning a hyper-plane near the diagonal axis designated by $x = y$, where x and y are the two numbers being compared. The fact that this hyper-plane is anchored at the origin and drifts away from the ideal diagonal at the higher values generates the min effect. The soft boundary created by the sigmoid activation function in CC networks produces the distance effect. In contrast, the rules learned by C4.5 made no psychological sense, e.g., *If $x > 5$ and $y > 7$, then $x > y$.*

Another coding trick employed by C4.5 modelers uses the difference between two numbers that are being compared (Schmidt & Ling, 1996). Mean proportions correct on training and test problems, respectively, were .902 and .875 for C4.5 difference coding in the interpolation experiment. This is an improvement, but again there is no coverage of the min and difference effects, and the rules are psychologically inappropriate, e.g., *If difference > 1, and $y > 2$, then $y > x$.*

In a study of extrapolation, the models were trained on pairs of the integers 0-4 and tested on pairs of the integers 5-9. There is no variation in C4.5 performance here because training patterns are not randomly selected for each run. Training and test results are shown in Table 1. Again, the CC algorithm learns and generalizes better than the C4.5 algorithm, whether input coding uses standard raw integers or differences.

Table 1: Proportion correct and generalization ratio for extrapolation.

Algorithm/coding	Train	Test	Ratio
CC	1.00	.99	.99
C4.5/standard	.56	.40	.71
C4.5/difference	.76	.40	.53

In conclusion, C4.5 does not learn or generalize well with either standard or difference coding of input on number comparison problems. It also fails to cover the min and difference effects, and the rules it learns are psychologically implausible. The only apparent way to get C4.5 to learn appropriate number comparison rules and generalize effectively is to build those rule conditions into the input coding, in which case there is nothing to learn. In contrast, CC learns and generalizes

well, while covering min and difference effects and generating reasonable knowledge representations, and it does so with raw numerical inputs.

Infant Familiarization to Sentences

The third case study concerns infant familiarization to sentences in an artificial language. A recent paper in this area has been of particular interest because it claimed to have data that could only be accounted for by rules and variables (Marcus, Vijayan, Rao, & Vishton, 1999). That study found that 7-month-olds attend longer to sentences with unfamiliar structures than to sentences with familiar structures. Particular features of the experimental design and some unsuccessful neural network models allowed the authors to conclude that unstructured neural networks cannot simulate these results. Several unstructured connectionist models have since disproved that claim (Shultz & Bale, 2001), but the current focus is on generalization ability of connectionist and rule-based models that learn representations of these sentences.

The present simulations focus in particular on Experiment 1 of Marcus et al. (1999). In this experiment, infants were familiarized to sentences with an ABA pattern, for example, *ga ti ga* or *li na li*. There were 16 of these ABA sentences, created by combining four A words (*ga*, *li*, *ni*, and *ta*) and four B words (*ti*, *na*, *gi*, and *la*). Subsequently, the infants were tested with two novel sentences that were consistent with the ABA pattern (*wo fe wo*, and *de ko de*) and two others that were inconsistent with ABA in that they followed an ABB pattern (*wo fe fe*, and *de ko ko*). There was also a condition in which infants were familiarized instead to sentences with an ABB pattern. Here the novel ABB sentences were consistent and the novel ABA sentences were inconsistent with the familiarized pattern. Infants attended more to inconsistent than to consistent novel sentences, suggesting that they were sensitive to syntactic properties of the sentences.

For consistency, I focus on a particular CC model of these data (Shultz & Bale, 2001). In this model, sentences were coded by real numbers representing the sonority (vowel likeness) of particular consonants or vowels. An encoder version of CC was used, enabling the network to learn to reproduce its inputs on its output units. Deciding on whether a particular sentence is correctly rendered in such networks is somewhat arbitrary. A more natural index of performance on training and test sentences is mean error, which is plotted in Figure 1. Test patterns inside the range of the training patterns were the same as those used with infants. Two additional sets tested extrapolation by using sonority values outside of the training range, by a distance that was either close or far. The greater error to inconsistent sentences corresponds to the attention difference found with infants. The fact that this consistency effect extends to patterns outside of the training range reveals substantial extrapolation ability in these networks. As well, the CC networks exhibited the typical exponential decrease in attention to familiarization stimuli that are found with infants.

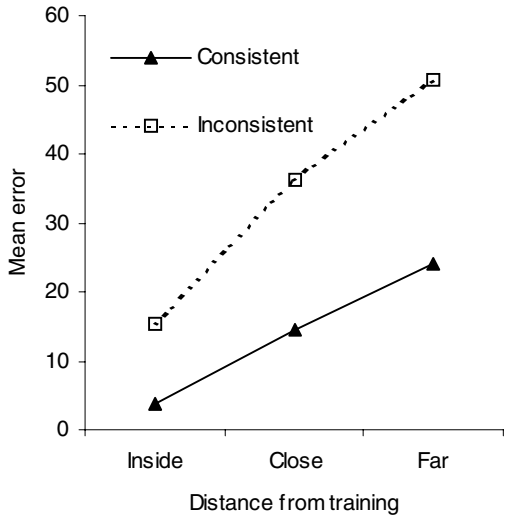


Figure 1: Mean error for CC networks simulating infant interest in consistent and inconsistent test sentences.

I did C4.5 simulations in several different ways to try to achieve successful learning and generalization. The initial attempt involved a literal symbolic encoding of each word in the sentences. For example, the word *ga* was coded as the symbol *ga*. Because there was only one output class when only one type of sentence was used as in the infant experiment (ABA or ABB), the resulting decision tree had only one leaf labeled with the syntactic class. In other words, if exposed only to ABA sentences, then expect more of the same. This is not really a rule and it captures none of the gradual characteristics of familiarization in infants. There is no variation in any of these C4.5 runs of the familiarization problem because each run uses all of the examples, rather than a random selection of examples.

The next C4.5 simulation added the 16 ABB sentences to the examples to be classified, in order to ensure that rules would be learned. This effectively changes the experiment to one of discrimination rather than familiarization. In this case, C4.5 focused only on the third word, concluding that the ABA syntax would be signaled by *ga*, *li*, *ni*, or *ta* as the third word, whereas the ABB syntax would be identified by *ti*, *na*, *gi*, or *la* as the third word. This is a sensible focus because the third word does distinguish the two syntactic types, producing a training success rate of 1.0, but it does not reflect Marcus et al.'s (1999) assumptions about infants

comparing the first and third words in each sentence. Moreover, because the test words are novel, this solution does not enable distinction between consistent and inconsistent test sentences. The generalization success rate is 0, as is the generalization ratio.

To obtain successful generalization with this kind of literal symbolic coding in C4.5, it is necessary to code the input relationally, explicitly representing equality of the first and second words, the first and third words, and the second and third words. When the first and third words are the same, then one has an ABA sentence; when the second and third words are the same then one has an ABB sentence. This allows perfect generalization to novel words, but the problem is that C4.5 can learn this relation perfectly with only one example of each pattern because the entire solution is explicitly presented in the inputs. Infants presumably require more examples than that to distinguish these syntactic patterns, reflecting the fact that their inputs are not coded so explicitly and fortuitously.

C4.5 was also trained with discrimination examples coded on sonority values as in the CC model. This model yielded 62.5% of training sentences correct, 0% correct on ABA and ABB test sentences, and a generalization ratio of 0. Moreover, the rules learned by this model were rather odd, e.g., *If C1 < -5, C3 < -5, and C2 > -6, then syntax is ABA*, where *C1* refers to the consonant of the first word, *C3* is the consonant of the third word, etc.

In contrast, the knowledge representations learned by the CC model were psychologically interesting. The hidden units were found to use sonority sums of the consonant and vowel to represent variation in sonority. This was achieved first in the duplicated-word category and next in the single-word category. This hidden unit representation was then decoded with similar weights to outputs representing the duplicate-word category.

Summarizing the results of the familiarization simulations, C4.5 did not show gradual familiarization effects. When the problem was changed to a discrimination problem, C4.5 did not learn the proper rules and did not generalize effectively. With explicit relational coding, C4.5 learns and generalizes perfectly, but it requires only two examples. When trained with sonority codes, C4.5 does not master the training examples, learns inappropriate rules, and does not generalize. In contrast, CC learns and generalizes well, both inside and outside of the range of the training examples, and acquires sensible knowledge representations.

Discussion

When learning of knowledge representations is required, CC reveals a number of advantages over C4.5: familiarizing to a single category, learning both simple (number comparison) and difficult (conservation) problems, finding structural relations that exist implicitly within training examples, learning rule-like functions that are psychologically plausible, covering psychological effects, and generalizing to novel examples, even to the extent of extrapolating outside of the training range. A pure comparison of generalization is difficult because of differences in learning success. However, comparison of generalization ratios that scale test performance by training performance, to control for learning success, consistently showed an advantage for CC over C4.5. This advantage occurred both with identical input coding for the two algorithms and with a variety of coding modifications that made it easier for C4.5 to learn.

Some of the generalization success of CC networks can be traced to the use of analog coding of inputs. In analog codes, the amount of unit activation varies with the intensity of the property being represented. Analog codes are well known to facilitate learning and generalization in artificial neural networks (Jackson, 1997), and exploratory comparative simulations suggest that they were important determinants of the present results. Their use in some of the present simulations can be justified by psychological evidence that people also employ analog representations, for example, of number.

Analog coding is not the entire story, however, because of two considerations. One is that not all of the CC inputs were analog. Some of the inputs to conservation problems that are essential to mature knowledge representations are coded in a discrete binary fashion. A second qualifier is that analog input codes were insufficient to allow successful learning and generalization in C4.5 models, even though C4.5 is equipped to deal with continuous inputs.

For a learning system to generalize effectively, it must of course learn the right sort of knowledge representation. This is why the present results show a close correspondence between success on the training examples and generalization performance. It was typical for performance to be slightly worse on test problems than on training problems, although generalization was considerably worse in some C4.5 runs, as indicated by low generalization ratios.

Because connectionist models generalized better than rule-based models under the learning constraint in three different domains, the argument that rule-based models show superior generalization is highly suspect. However, it is reasonable to ask whether connectionist models invariably generalize better than rule-learning models. Would this finding hold up in different domains and with different learning algorithms? Obviously, more research is needed, but we are now beyond facile comparisons of hand-written or imagined rules to laboriously learned connections.

Choice of algorithm is a key issue because both symbolic and neural algorithms may vary considerably in their ability to learn and generalize. Certainly, CC benefits from its ability to learn difficult problems that are beyond the ability of other neural learning procedures and its tendency to build the smallest network necessary to master the problem on which it is being trained. Likewise, C4.5 benefits from its use of information gain to select the best feature on which to partition unclassified examples. Both algorithms have led the way in their respective class in producing successful simulations of cognitive development. Nonetheless, it is important for other algorithms of each type to be tried. It is possible that other rule-learning algorithms would have better success in finding more abstract and thus more general knowledge representations than C4.5 does. Although C4.5 is adept at learning from examples, it seems unable to represent those examples in anything more abstract than the features used in their input descriptions. This limitation could make learning and generalization difficult.

Finally, it is important to stress that generalization ability should not be taken as the ultimate criterion on which to evaluate different cognitive models. Surely, it is more critical to determine whether a given model generalizes like human subjects do. This is an issue that has not yet been adequately addressed.

Acknowledgments

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. Yoshio Takane, Alan Bale, and Francois Rivest contributed insightful comments on an earlier draft.

References

Anderson, J. A. (1995). *An introduction to neural networks*. Cambridge, MA: MIT Press.

Buckingham, D., & Shultz, T. R. (2000). The developmental course of distance, time, and velocity concepts: A generative connectionist model. *Journal of Cognition and Development, 1*, 305-345.

Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture. In D. S. Touretzky (Ed.), *Advances in neural information processing systems 2* (pp. 524-532). Los Altos, CA: Morgan Kaufmann.

Jackson, T. O. (1997). Data input and output representations. In E. Fiesler & R. Beale (Eds.), *Handbook of neural computation*. Oxford: Oxford University Press.

Klahr, D. (1984). Transition processes in quantitative development. In R. J. Sternberg (Ed.), *Mechanisms of cognitive development*. New York: Freeman.

Ling, C. X. (1994). Learning the past tense of English verbs: The symbolic pattern associator vs. connectionist models. *Journal of Artificial Intelligence Research, 1*, 209-229.

Ling, C. X., & Marinov, M. (1993). Answering the connectionist challenge: A symbolic model of learning the past tenses of English verbs. *Cognition, 49*, 235-290.

Ling, C. X., & Marinov, M. (1994). A symbolic model of the nonconscious acquisition of information. *Cognitive Science, 18*, 595-621.

Ling, C. X., & Wang, H. (1996). *A decision-tree model for reading aloud with automatic alignment and grapheme generation*. Unpublished paper, Department of Computer Science, University of Western Ontario.

Marcus, G. (1998). Rethinking eliminative connectionism. *Cognitive Psychology, 37*, 243-282.

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science, 283*, 77-80.

Mareschal, D., & Shultz, T. R. (1999). Development of children's seriation: A connectionist approach. *Connection Science*, *11*, 149-186.

Oshima-Takane, Y., Takane, Y., & Shultz, T. R. (1999). The learning of first and second pronouns in English: Network models and analysis. *Journal of Child Language*, *26*, 545-575.

Pinker, S. (1997). *How the mind works*. New York: Norton.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*, 81-106.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Reed, R., & Marks II, R. J. (1988). Neurosmithing: Improving neural network learning. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks*. Cambridge, MA: MIT Press.

Schmidt, W. C., & Ling, C. X. (1996). A decision-tree model of balance scale development. *Machine Learning, 24*, 203-229.

Shastri, L. (1995). Structured connectionist models. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks*. MIT Press, Cambridge, MA.

Shultz, T. R. (1998). A computational analysis of conservation. *Developmental Science*, *1*, 103-126.

Shultz, T. R., & Bale, A. C. (2001, in press). Neural network simulation of infant familiarization to artificial sentences: Rule-like behavior without explicit rules and variables. *Infancy*.

Shultz, T. R., Mareschal, D., & Schmidt, W. C. (1994). Modeling cognitive development on balance scale phenomena. *Machine Learning*, *16*, 57-86.

Sirois, S., & Shultz, T. R. (1998). Neural network modeling of developmental effects in discrimination shifts. *Journal of Experimental Child Psychology, 71*, 235-274.

Clinging to Beliefs: A Constraint-satisfaction Model

Thomas R. Shultz (shultz@psych.mcgill.ca)
Department of Psychology; McGill University
Montreal, QC H3C 1B1 Canada

Jacques A. Katz (jakatz@cnbc.cmu.edu)
Department of Psychology; Carnegie Mellon University
Pittsburgh, PA 15213 USA

Mark R. Lepper (lepper@psych.stanford.edu)
Department of Psychology; Stanford University
Stanford, CA 94305-2130 USA

Abstract

Beliefs tend to persevere even after evidence for their initial formulation has been invalidated by new evidence. If people are assumed to rationally base their beliefs on evidence, then this belief perseverance is somewhat counterintuitive. We constructed a constraint-satisfaction neural network model to simulate key belief perseverance phenomena and to test the hypothesis that explanation plays a central role in preserving evidentially challenged beliefs. The model provides a good fit to important psychological data and supports the hypothesis that explanations preserve beliefs.

Introduction

It is perhaps surprising that people are so often reluctant to abandon personal beliefs that are directly contradicted by new evidence. This tendency to cling to beliefs in the face of subsequent counterevidence has been well demonstrated for opinions (Abelson, 1959), decisions (Janis, 1968), impressions of people (Jones & Goethals, 1971), social stereotypes (Katz, 1960), scientific hypotheses (T. S. Kuhn, 1962), and commonsense ideas (Gilovich, 1991).

Belief perseverance is puzzling because it is commonly assumed that beliefs are based on evidence. If it is rational for people to form a belief based on evidence, then why is it not equally rational for them to modify the belief when confronted with evidence that invalidates the original evidence?

Debriefing Experiments

Some of the clearest cases of apparently irrational belief perseverance come from debriefing experiments. In these experiments, subjects learn that the initial evidential basis for a belief is invalid. For example, Ross, Lepper, and Hubbard (1975) first provided subjects with false feedback concerning their ability to perform a novel task. Their subject's task was to distinguish authentic from fake suicide notes by reading a number of examples. False feedback from the experimenter led subjects to believe that they had performed at a level that was much better than average or much worse than average. Then, in a second phase, subjects were debriefed about the random and predetermined nature of the feedback that they had received in the first phase. There were three debriefing conditions. In the outcome debriefing condition, subjects were told that the evidence on which their initial beliefs were based had been completely fabricated by the experimenter. Subjects in the process debriefing condition were additionally told about the procedures of outcome debriefing along with explanations about possible mechanisms and results of belief perseverance. Subjects in this condition were also told that belief perseverance was the focus of the experiment. Finally, subjects in a no-debriefing control condition were not debriefed at all after the feedback phase. Subsequently, subjects in all three conditions rated their own ability at the suicide-note verification task. This was to assess the perseverance of their beliefs about their abilities on this task that were formed in the feedback phase.

The mean reported beliefs for the three debriefing conditions are shown in Figure 1. There is an interaction between debriefing condition and the nature of feedback (success or failure at the note-verification task). The largest difference between success and failure feedback occurs in the no-debriefing condition. In this control condition, subjects who were initially led to believe that they had succeeded continue to believe that they would do better than subjects initially led to believe that they had failed. After outcome debriefing, there is still a significant difference between the success and failure conditions, but at about one-half of the strength of the control condition. The difference between success and failure feedback effectively disappears after process debriefing. This sort of belief perseverance after debriefing has been convincingly demonstrated for a variety of different beliefs and debriefing techniques (Jennings, Lepper, & Ross, 1981; Lepper, Ross & Lau, 1986).

One explanation for such belief perseverance is that people frequently explain events, including their own beliefs, and such explanations later sustain these beliefs in the face of subsequent evidential challenges (Ross et al., 1975). For example, a person who concludes from initial feedback that she is very poor at authenticating suicide notes might attribute this inability to something about her experience or personality. Perhaps she has had too little contact with severely depressed people, or maybe she is too optimistic to empathize deeply with a suicidal person. Then in the second

phase, when told that the feedback was entirely bogus, these previously constructed explanations may still suggest that she lacks the ability to authenticate suicide notes. Analogously, a subject who is initially told that he did extremely well at this task may explain his success by noting his familiarity with some depressed friends or his sensitivity to other people's emotions. Once in place, such explanations could inoculate the subject against subsequent evidence that the initial feedback was entirely bogus.

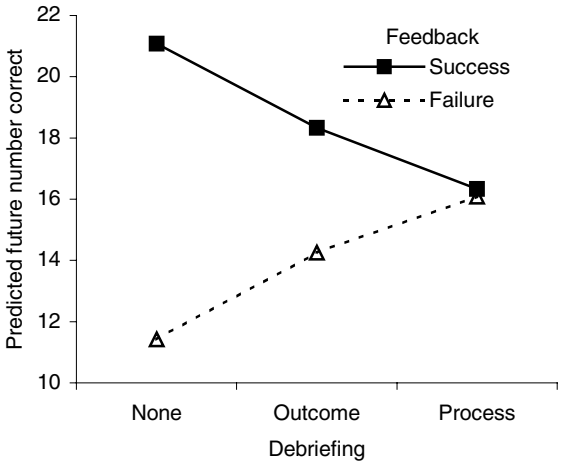


Figure 1: Mean predicted ability in the Ross et al. (1975) experiment after debriefing.

The assumption is that even though contradictory evidence may weaken a belief, it is unlikely to alter every cognition that may have derived from that belief, such as explanations for the belief's existence. The well-known frame problem emphasizes the computational intractability of tracking down every implication of an altered belief (Charniak & McDermott, 1985). People generally do not have the time, energy, knowledge, or inclination to decide which other beliefs to change whenever a belief is changed.

In contrast to the view that people have difficulty distinguishing explanations from evidence (D. Kuhn, 1991), recent research suggests that people can distinguish explanations from evidence and that they tend to use explanations as a substitute for missing evidence (Brem & Rips, 2000).

In this paper, we report on our attempt to simulate the belief perseverance phenomena reported by Ross et al. (1975). Our basic theoretical premise in designing these simulations is that belief perseverance is a special case of a more general tendency for people to seek cognitive consistency. Striving for consistency has long been considered to cause a wide variety of phenomena in social psychology (Abelson, Aronson, McGuire, Newcomb, Rosenberg, & Tannenbaum, 1968). In the case of belief perseverance, we assume that people form percepts that are consistent with external evidence, then acquire beliefs that are consistent with these percepts, and finally construct explanations that are consistent with these beliefs. We view resistance to new evidence that contradicts existing percepts, beliefs, or explanations as part of an attempt to achieve overall consistency among current cognitions, given that not all implications of contradictory evidence are actively pursued.

There was a simulation using non-monotonic logic of how belief can be preserved despite ordinary debriefing, but it did not cover the quantitative differences between conditions in the Ross et al. experiment (Hoenkamp, 1987).

Neural Constraint Satisfaction

Our simulations use a technique called constraint satisfaction, which attempts to satisfy as many constraints as well as possible within artificial neural networks. The present model is closely related to models used in the simulation of schema completion (Rumelhart, Smolensky, McClelland, & Hinton, 1986), person perception (Kunda & Thagard, 1996), attitude change (Spellman, Ullman, & Holyoak, 1993), and dissonance reduction (Shultz & Lepper, 1996).

Constraint satisfaction neural networks are comprised of units connected by weighted links. Units can represent cognitions by taking on activation values from 0 to 1, representing the strength or truth of the cognition. Connection weights can represent relations between cognitions and are assigned positive or negative values representing the sign and strength of the relations. Connection weights are bidirectional to permit cognitions to mutually influence each other. External inputs to units represent influences from the environment. Biases are represented by internal inputs to a given unit that do not vary across different network inputs.

Networks attempt to satisfy the soft constraints imposed by fixed inputs, biases, and weights by changing activation values of the units. Unit activations are updated according to these rules:

$$a_i(t+1) = a_i(t) + net_i\left(ceiling - a_i(t)\right), \text{ when } net_i \geq 0 \quad (1)$$

$$a_i(t+1) = a_i(t) + net_i\left(a_i(t) - floor\right), \text{ when } net_i < 0 \quad (2)$$

where $a_i(t + 1)$ is the updated activation value of unit i, a_i is the current activation of unit i, ceiling is the maximum activation value for a unit, floor is the minimum activation value for a unit, and net_i is the net input to unit i, as computed by:

$$net_i = in\left(\sum_j w_{ij} a_j + bias_i\right) + ex(input_i) \quad (3)$$

where in and ex are parameters that modulate the impact of the internal and external inputs, respectively, with default values of 0.1, w_{ij} is the connection weight between units i and j, a_j is the activation of sending unit j, $bias_i$ is the bias value of unit i, and $input_i$ is the external input to unit i.

These update rules ensure that network consistency either increases or stays the same, where consistency is computed as:

$$consistency = \sum_{ij} w_{ij} a_i a_j + \sum_i input_i a_i + \sum_i bias_i a_i \quad (4)$$

When a network reaches a high level of consistency, this means that it has settled into a stable pattern of activation and that the various constraints are well satisfied. In such stable solutions, any two units connected by positive weights tend to both be active, units connected by negative weights tend not to be simultaneously active, units with high inputs tend to be more active than units with low inputs, and units with high biases tend to be more active than units with low biases.

Increases in consistency and constraint satisfaction occur gradually over time. At each time cycle, *n* units are randomly selected for updating, where *n* is typically the number of units in the network. Thus, not every unit is necessarily updated on every cycle and some units may be updated more than once on a given cycle.

Unusual Simulation Features

The foregoing characteristics of neural constraint satisfaction are quite common. In addition, the present modeling has a few somewhat unusual features. Perhaps the most important of these is a two-phase structure that accommodates the two main phases of belief perseverance experiments. It is more typical for neural constraint satisfaction models to operate in a single phase in which networks are designed and updated until they settle into a stable state. Our two phases correspond to the feedback and debriefing phases of these experiments. After a network settles in the initial feedback phase, new units can be introduced, and inputs, connection weights, and biases may be changed in a second, debriefing phase. To implement continuity between the two phases, a simple type of memory was introduced such that activation values from the feedback phase would be partially retained as unit biases in the debriefing phase. Final activations in the feedback phase were multiplied by 0.05 to transform them into biases for the debriefing phase. This is not a detailed implementation of a memory model, but is rather a convenient shorthand implementation of the idea that there is a faded memory for whatever conclusions were reached in the previous, feedback phase.

Two other unusual features derived from our earlier simulations of cognitive dissonance reduction (Shultz & Lepper, 1996): a cap parameter and randomization of network parameters. The cap parameter is a negative self-connection weight for every unit that limits unit activations to less than extreme values. The purpose of this activation cap is to increase psychological realism for experiments about beliefs that reach no more than moderate strength.

Robustness of simulation results was assessed by simultaneously randomizing all network parameters (i.e., biases, inputs, and connection weights) by up to 0%, 10%, 50%, or 100% of their initial values according the formula:

$$y = x \pm \{rand \ (abs \lfloor x * rand\% \rfloor) \} \quad (5)$$

The initial parameter value x is multiplied by the proportion of randomization being used (0, .1, .5, or 1) and converted to an absolute value. Then a random number is selected between 0 and the absolute value under a uniform distribution. This random number is then randomly either added to or subtracted from the initial value. This parameter randomization allows efficient assessment of the robustness of the simulation under systematic variations of parameter values. If the simulations succeed in matching the psychological data, even under high levels of parameter randomization, then they do not depend on precise parameter settings. This randomization process also enhances psychological realism because not every subject can be expected to have precisely the same parameter values.

Network Design

Units

Units represent external input and the three types of cognitions that are critical to belief perseverance experiments, i.e., percepts, beliefs, and explanations. Percept units represent a subject's perception of external input, in this case feedback provided by the experimenter. Belief units represent a subject's beliefs, and explanation units represent a subject's explanations of particular beliefs. In each case, the larger the activation value of a given unit, the stronger the associated cognition. Activation values range from 0 to 1, with 0 representing no cognition, and 1 representing the strongest cognition. All unit activations start at 0 as a network begins to run.

Unit names include a sign of +, -, or 0 to represent the direction of a given cognition. For example, in these simulations, *+percept* refers to a perception of doing well on a task, *-percept* to a perception of doing poorly on the task, and *0percept* to not knowing about performance on the task. Percept units sometimes have an external input, to reflect the feedback on which the percept is based. A *0percept* unit is required for simulating debriefing experiments, where information is encountered that explicitly conveys a lack of knowledge about performance. Analogously, *+belief* represents a belief that one is performing well at a task, *–belief* represents a belief that one is performing poorly at a task, *+explanation* represents an explanation for a *+belief,* and *-explanation* represents an explanation for a *–belief*.

Connections

Units are joined by connection weights that have a size and a sign. The sign of a weight represents a positive or negative relation between connected units. A positive weight signals that a cognition follows from, leads to, is in accordance with, or derives support from another cognition. A negative weight indicates that a cognition is inconsistent with or interferes with another cognition. Decisions about signs are based on descriptions of psychological procedures. Initial nonzero connection weights are + or - 0.5 in our simulations. Connection weights of 0 indicate the absence of relations between cognitions. All connection weights are bi-directional to allow mutual influences between cognitions.

The general connection scheme in our simulations of belief perseverance has external inputs feeding percepts, which are in turn connected to beliefs, which are in turn connected to explanations. For failure conditions, a -percept unit receives external input and is connected to a -belief unit, which is in turn connected to a -explanation unit. For success conditions, a +percept unit receives external input and is connected to a +belief unit, which is in turn connected to a +explanation unit. Connection weights between incompatible cognitions, such as between +belief and -belief or between -percept and 0percept, are negative.

The principal dependent measure in many belief perseverance studies is a subject's self-rated ability on a task. This is represented as net belief, computed as activation on the +belief unit minus activation on the -belief unit, after the network settles in the debriefing phase. This technique of using two negatively connected units to represent the different poles of a single cognition was used by Shultz and Lepper (1996) in their simulation of cognitive dissonance phenomena.

Networks for Feedback Phase

Figure 2 shows specifications for the negative feedback condition. Negative feedback, in the form of external input, with a value of 1.0, is positively connected to the –percept unit. This same network design is used for the no-debriefing condition of the debriefing phase.

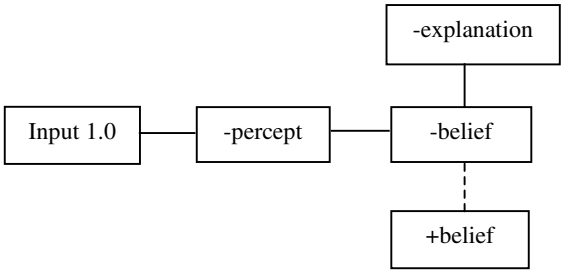


Figure 2: Network for negative feedback. Positive connection weights are indicated by solid lines; negative connection weights by dashed lines.

A feedback phase represents the presentation of information on how a subject is doing on a task. It is assumed that this information forms the basis for a belief about ability and to explanations of that ability. Because of the connection scheme and the fact that all unit activations start at 0, percept units reach activation asymptotes first, followed by belief units, and finally by explanation units.

Networks for Debriefing Phase

Figure 3 shows network specifications for the debriefing phase. This network was used for both outcome debriefing and process debriefing. The particular network shown in Figure 3 shows a debriefing phase that follows negative feedback. As noted earlier, an unusual feature here is the inclusion of biases for percept, belief, and explanation units from the earlier, feedback phase. These biased units are represented by bolded rectangles around unit names, and implement a faded memory of the feedback phase. There is also a new unit, the 0percept unit, with an input of 1.0, to represent that nothing valid is known about task performance. This unit has no bias because it was not present in the previous phase. It is negatively connected to the – or + percept unit to represent the idea that the feedback data from the previous phase are false, and thus convey no information about task ability.

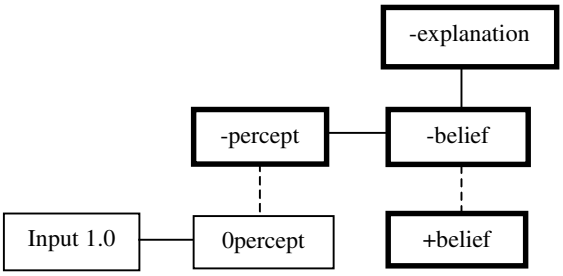


Figure 3: Network for outcome and process debriefing following negative feedback. Units that have biases from the feedback phase are indicated by bolded rectangles.

We implemented the stronger, process debriefing by multiplying bias values by a factor of 0.1. This reflects the idea that process debriefing is so thorough that it severely degrades all cognitions that were created in the preceding feedback phase. Networks in the no-debriefing condition were identical to those described in Figure 2, with no topology changes after the feedback phase. However, as in all debriefing conditions, biases of .05 of final activations were used for any units being carried over from the feedback phase. Networks were run for 120 update cycles in each of the two phases; by this time they had typically settled into stable states.

Principles of Network Design

In summary, network design can be summarized by 13 principles:

1. Units represent cognitions.
2. The principal cognitions in belief perseverance experiments are input feedback, percepts, beliefs, and explanations.
3. The sign of unit names represent the positive or negative poles of cognitions.
4. Unit activation represents strength of a cognition (or a pole of a cognition).
5. The difference between positive and negative poles of a cognition represents the net strength of the cognition.
6. Connection weights represent constant implications between cognitions.
7. Connection weights are bi-directional, allowing possible mutual influence between cognitions or poles of cognitions.

8. Cognitions whose poles are mutually exclusive have negative connections between the positive and negative poles.
9. Size of external input represents strength of environmental influence, such as evidence or feedback.
10. External inputs are connected to percepts, percepts to beliefs, and beliefs to explanations, representing the assumed chain of causation in belief perseverance experiments. That is, environmental feedback creates percepts, which in turn create beliefs, which eventually lead to explanations for the beliefs.
11. Networks settling into stable states represent a person's tendency to achieve consistency among cognitions.
12. Final unit activations from the feedback phase are converted to unit biases for the start of the debriefing phase of belief perseverance experiments, representing the participant's memory of the feedback phase.
13. Multiplying activation bias values by 0.1 represents thorough, process debriefing.

Results

We focus on the final net belief about one's ability after the debriefing phase. This is computed as activation on the +belief unit minus activation on the -belief unit. Here, we report only on the 10% randomization level, but similar results are found at each level of parameter randomization.

Net belief scores were subjected to a factorial ANOVA in which debriefing condition (none, outcome, and process) and feedback condition (success, failure) served as factors. There was a main effect of feedback, $F(1, 114) = 29619$, $p < .001$, and an interaction between debriefing and feedback, $F(2, 114) = 9102$, $p < .001$. Mean net ability scores are shown in Figure 4. For success feedback, net belief scores were higher after no debriefing than scores obtained after outcome debriefing, which were in turn higher than scores obtained after process debriefing. The opposite holds for failure feedback.

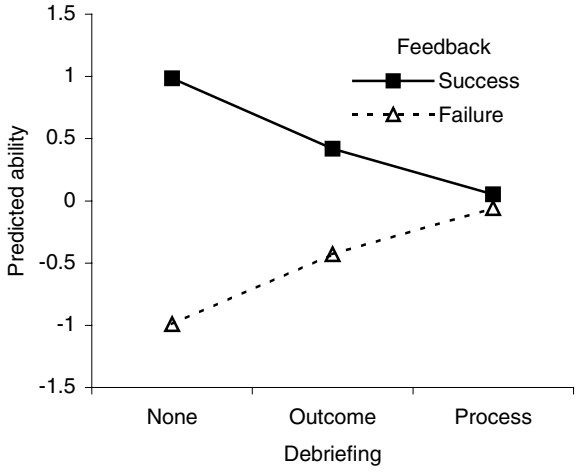


Figure 4: Mean predicted ability in the simulation after debriefing.

To assess the fit to human data, we computed a regression F with regression weights based on the pattern of the Ross et al. (1975) results. The regression weights were 2, -2, 1, -1, 0, and 0 for the no debriefing/success, no debriefing/failure, outcome debriefing/success, outcome debriefing/failure, process debriefing/success, and process debriefing/failure conditions, respectively. This produced a highly significant regression $F(1, 114) = 47558$, $p < .001$, with a much smaller residual $F(4, 114) = 67$, $p < .001$. The regression F accounts for 99% of the total variance in net belief. As with human subjects, there is a large difference between success and failure with no debriefing, a smaller but still substantial difference after outcome debriefing, and very little difference after process debriefing.

To assess the role of explanation in the simulation, we subjected activations on the explanation unit after the debriefing phase to the same ANOVA. There is a main effect of debriefing, $F(2, 114) = 3787$, $p < .001$, a much smaller main effect for feedback $F(1, 114) = 15.37$, $p < .001$, and a small interaction between them, $F(2, 114) = 6.76$, $p < .005$. The mean explanation scores are presented in Figure 5. Explanations are strong under no-debriefing, moderately strong under outcome debriefing, and weak under process debriefing. But because explanations had been strongly active in all three conditions at the end of the feedback phase, these post-debriefing results reflect relative differences in maintenance of explanations. Explanations are maintained under no debriefing, partially maintained under outcome debriefing, and eliminated in process debriefing.

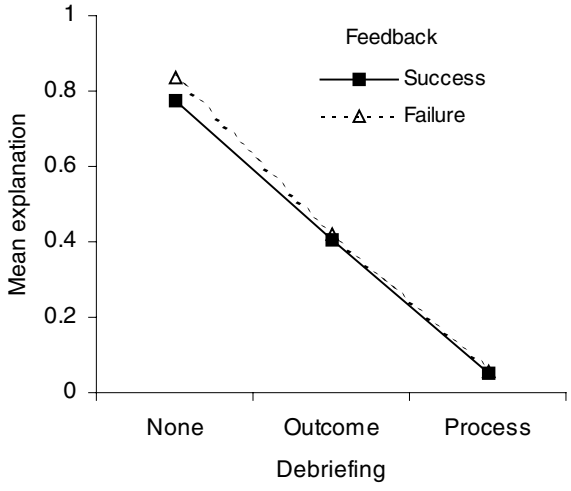


Figure 5: Mean explanation scores in the simulation after debriefing.

Discussion

The tendency for beliefs to persevere even after evidence for them has been fully invalidated challenges some basic assumptions about human rationality. If people reasonably

base their beliefs on evidence, then why is counter-evidence not sufficient to eliminate or change beliefs?

We used constraint-satisfaction neural networks to test the idea that explanation plays a key role in sustaining beliefs in these circumstances. The model provides a good fit to existing psychological data from debriefing experiments in which subjects are informed that that the principal evidence for their beliefs is no longer valid (Ross et al., 1975). Simulated beliefs remain strong without debriefing; belief strength is reduced after standard outcome debriefing, and eliminated after more thorough, process debriefing. This pattern of results matches the psychological data, with about half-strength beliefs under outcome debriefing and elimination of beliefs by process debriefing. As in our earlier simulations of cognitive dissonance phenomena, the neural constraint-satisfaction model is here shown to be robust against parameter variation. Even a high degree of parameter randomization does not change the pattern of results.

The simulations further revealed that belief perseverance is mirrored by strength of explanation. Explanations remain strong with no debriefing, and decrease progressively with more effective debriefing. Although it is obvious that debriefing reduces the strength of erroneous beliefs, the finding that it also reduces explanations is perhaps less obvious. In our simulations, explanation is reduced by effective debriefing via connections from external evidence to percepts, percepts to beliefs, and beliefs to explanations.

People spontaneously generate explanations for events as a way of understanding events, including their own beliefs (Kelley, 1967). If an explanation is generated, this explanation becomes a reason for holding an explained belief, even if the belief is eventually undercut by new evidence.

Future work in our group will extend this model to other belief perseverance phenomena and attempt to generate predictions to guide additional psychological research.

Acknowledgments

This research was supported by a grant to the first author from the Social Sciences and Humanities Research Council of Canada and by grant MH-44321 to the third author from the U.S. National Institute of Mental Health.

References

Abelson, R. P. (1959). Modes of resolution of belief dilemmas. *Conflict Resolution, 3*, 343-352.

Abelson, R. P., Aronson, E., McGuire, W. J., Newcomb, T. M., Rosenberg, M. J., & Tannenbaum, P. H. (Eds.) (1968). *Theories of cognitive consistency: A sourcebook*. Chicago: Rand McNally.

Brem, S. K., & Rips, L. J. (2000). Explanation and evidence in informal argument. *Cognitive Science, 24*, 573-604.

Charniak, E., & McDermott, D. (1985). *Introduction to artificial intelligence*. Reading, MA: Addison-Wesley.

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life*. New York: Free Press.

Hoenkamp, E. (1987). An analysis of psychological experiments on non-monotonic reasoning. *Proceedings of the Tenth International Joint Conference on Artificial Intelligence* (Vol. 1, pp. 115-117). Los Altos, CA: Morgan Kaufmann.

Janis, I. (1968). Stages in the decision-making process. In R. P. Abelson, E. Aronson, W. J. McGuire, T. M. Newcomb, M. J. Rosenberg, & P. H. Tannenbaum (Eds.) (1968). *Theories of cognitive consistency: A sourcebook*. Chicago: Rand McNally.

Jennings, D. L., Lepper, M. R., & Ross, L. (1981). Persistence of impressions of personal persuasiveness: Perseverance of erroneous self-assessments outside the debriefing paradigm. *Personality and Social Psychology Bulletin, 7*, 257-263.

Jones, E. E., & Goethals, G. R. (1971). Order effects in impression formation: Attribution context and the nature of the entity. In E. E. Jones et al. (Eds.), *Attribution: Perceiving the causes of behavior*. Morristown, NJ: General Learning Press.

Katz, D. (1960). The functional approach to the study of attitudes. *Public Opinion Quarterly, 24*, 163-204.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska Symposium on Motivation.* Vol. 15. Lincoln: University of Nebraska Press.

Kuhn, D. (1991). *The skills of argument*. Cambridge: Cambridge University Press.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint satisfaction theory. *Psychological Review, 103*, 284-308.

Lepper, M. R., Ross, L., & Lau, R. R. (1986). Persistence of inaccurate beliefs about self: Perseverance effects in the classroom. *Journal of Personality and Social Psychology, 50*, 482-491.

Ross, L, Lepper, M. R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology, 32*, 880-892.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. (1986). Schemata and sequential thought processes in PDP models. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2). Cambridge, MA: MIT Press.

Shultz, T. R., & Lepper, M. R. (1996). Cognitive dissonance reduction as constraint satisfaction. *Psychological Review, 103*, 219-240.

Spellman, B. A., Ullman, J. B., & Holyoak, K. J. (1993). A coherence model of cognitive consistency: Dynamics of attitude change during the Persian Gulf war. *Journal of Social Issues, 49*, 147-165.

Semantic Effect on Episodic Associations

Yaron Silberman (yarons@alice.nc.huji.ac.il)
Interdisciplinary Center for Neural Computation, The Hebrew University of Jerusalem
Giv'at Ram, Jerusalem 91904 Israel

Risto Miikkulainen (risto@cs.utexas.edu)
Department of Computer Science, The University of Texas at Austin
Austin, TX 78712-1188 USA

Shlomo Bentin (msbentin@mscc.huji.ac.il)
Department of Psychology, The Hebrew University of Jerusalem
Mount Scopus; Jerusalem 91905 Israel

Abstract

We examined the influence of the pre-existing organization of the semantic memory on forming new episodic associations between words. Testing human subjects' performance we found that a semantic relationship between words facilitates forming episodic associations between them. Furthermore, the amount of facilitation increases linearly as a function of the number of co-occurrence of the words, up to a ceiling. Constrained by these empirical findings we developed a computational model, based on the theory of spreading activation over semantic networks. The model uses self-organizing maps to represent semantic relatedness, and lateral connections to represent the episodic associations. When two words are presented to the model, the interaction of the two activation waves is summed and added to the direct lateral connection between them. The main result is that the model is capable of replicating the empirical results. The model also makes several testable predictions: First, it should be easier to form an association from a word with few semantic neighbors to a word with many semantic neighbors than vice-versa. Second, after associating an unrelated word pair it should be easier to associate another two words each related to one of the words in the first pair. Third, a less focused activation wave, which may be the cause of schizophrenic thought disorder, should decrease the advantage in learning rate of related over unrelated pairs.

Introduction

The principles of forming associations between concepts in memory have been studied since the early days of psychological research. For example, the British empiricist school of philosophers (e.g., Hume, 1738), proposed three main principles of association: Contiguity (i.e., proximity in time and space), Similarity (or Contrast), and Cause and Effect. Fulfilling any of these conditions should be sufficient to form an association between concepts. The strength of the association is determined by the frequency at which any of the above conditions is fulfilled. An important aspect of this theory is that intentionality is not a necessary condition for the associative process to occur. Indeed, associations are frequently established without intention and without allocating attention to the learning process. We will refer to these associations as *incidental associations*. This paper presents a computational component of a larger study in which we examine characteristics of forming incidental associations between words.

Phenomenologically defined, two words are associated if the presentation of one brings the second to the perceiver's awareness. Associations between words can be formed in at least two different ways. First, *episodic associations* are formed when two words co-occur in time and space. An episodic association is therefore a subjective experience. Second, *semantic associations* are based on semantic relatedness between the words. Words are considered semantically related if they share common semantic features, for example, if they belong to the same semantic category. Although the two classes of associations are based on different properties, many associated word pairs are also semantically related, which raises the possibility of an interaction between the two types of associations.

A well-known interaction of this type is the semantic priming effect (Meyer and Schvaneveldt, 1971). The presentation of a related prime word prior to performing a lexical task, such as naming and/or lexical decision, results in faster and more accurate performance (see Neely, 1991, for a review). Ample research was aimed at isolating the types of word relations that mediate this phenomenon (e.g. Fischler, 1977; McKoon & Ratcliff, 1979; Moss et al., 1995; Shelton & Martin, 1992). More specifically, a frequently asked question was whether words that are related only in one way, either semantically or episodically, would induce effective priming and how an interaction between these to types of relations would affect priming. Although a debate still exists, it is safe to say that both types of relations prime effectively and that their combined effect is additive.

A common theory of the organization principles of the semantic system and the mechanisms underlying semantic priming is the theory of spreading activation over semantic networks (Collins & Loftus, 1975). In a semantic network, a concept is represented as a node. Semantically related nodes are connected with unidirectional weighted links. When a concept is processed,

the appropriate node is activated and the activation spreads along the connections of the network, gradually decreasing in strength. The decrease is proportional to the weights of the links in the path. In addition, the activation decays over time. Awareness of word occurs when its activation exceeds a threshold. According to this theory semantic priming occurs when activation from the prime spreads and partially pre-activates related targets so that a smaller amount of additional processing is required to bring its node activation above threshold.

An alternative and computationally more explicit modeling approach was recently proposed to explain semantic priming. Such models represent concepts in the semantic system by distributed (rather than local) representations. Concepts are not represented by single units, but rather by distinguishable patterns of activity over a large number of units (Hinton, 1990; Masson, 1995; Moss et al., 1994; Plaut, 1995). Each participating unit accounts for a specific semantic microfeature, and semantic similarity is thus expressed as overlap in activity patterns over the set of micro-features. In these models, recurrent dynamics is employed until the net settles to a stable state (an attractor). Semantic priming is explained by the fact that after settling to a prime, fewer modifications in the nodes' state are necessary for settling to a related target, making the latter settling process faster.

All the currently computational models of semantic priming have focused on processes based on existing associations. The process of acquiring new associations was abstracted in the training of the network. In the current study, we propose a computational model of forming new episodic associations between words on the basis of an already existing semantic network and show how this process is influenced by the organization of the semantic system.

Behavioral Experiments

A series of human performance experiments was conducted to supply constraints on the associating process. The following is a brief description of the relevant experiments and results (see Silberman & Bentin, submitted, for an elaborated report).

In one experiment, 10 randomly ordered Hebrew word pairs were repeated 20 times. In each trial, the two words were displayed one after the other with a Stimulus Onset Asynchronicity (SOA) of 700 ms. The subjects searched whether a letter presented 800 ms after the onset of the second word was included in the preceding word pair. Hence, proximity was achieved by having the subjects store the two words together in working memory for 800 ms. Following this "study session", the strength of the association between the words in each pair were unexpectedly tested using cued recall and a free association tests. In the cued recall test, the subjects were presented with the first words that occurred in half of the pairs, and asked to remember each word's associate. In the free association test, they were presented with the first words the other pairs, and asked to respond with their first free associate. We compared the strength of incidentally formed associations between semantically related (e.g. *milk-soup*) and semantically unrelated words (e.g. *paint-forest*). The results of this experiment, based on 64 subjects, are presented in Table 1.

Table 1: Percentage of cued recall and free association for pairs of semantically related and unrelated words.

Relatedness	Cued Recall	Free Associations
Related	38.8%	7.5%
Unrelated	19.4%	1.3%[1]

[1] Based on 16 subjects only.

As is evident in Table 1, semantic relatedness between words doubled the probability that an association would be incidentally formed between them. A between-subjects ANOVA of the cued recall performance showed that the difference between the two groups was statistically reliable [$F(1,62)=7.84$, $p<0.01$].

If semantic relationship facilitates the formation of associations by providing a higher initial linkage baseline or a smaller pool of candidates in the test phase, its effect should not interact with the number of episodic repetitions. Hence the difference between recall performance for related and unrelated word pairs should be the same, regardless the number of repetitions in the incidental learning phase (obviously, the absolute performance for both groups should positively correlate with the number of repetitions, up to a ceiling effect). To test this hypothesis, we manipulated the number of times each pair of the semantically related and unrelated pairs was repeated during the study phase.

Twenty-four Hebrew word pairs were selected for this experiment. The words in each pair were semantically related (belonged to the same semantic category) but not strongly associated (verified using free association questionnaires, in which we tested that none of the words was elicited by its pair among the first three free associates). Two study lists were prepared. Each consisted of 12 originally related pairs and 12 unrelated pairs formed by randomly pairing the other words. Pairs presented in the related condition in one list were used to form the pairs of the unrelated condition in the other list. Four groups of 24 subjects each were assigned to either 1 presentation (i.e., no repetition), 5, 10 or 20 presentations during the incidental study phase, in which subjects performed in the letter search task. The results of this experiment are presented in Figure 1.

An ANOVA showed that semantically related pairs were associated better than semantically unrelated pairs [$F(1,92)=204$, $p<0.0001$], and that the main effect of the number of repetitions was significant [$F(3,92)=25$ $p<0.0001$]. More revealing, however, was the significant interaction between the two factors [$F(3,92)=19$, $p<0.0001$], suggesting that each repetition contributed

more to related than to unrelated pairs. These results suggest that semantic information reinforces the formation of associations (at least if these are formed incidentally). The semantic effect has a ceiling at which additional repetition contributes equally to forming both related and unrelated associations.

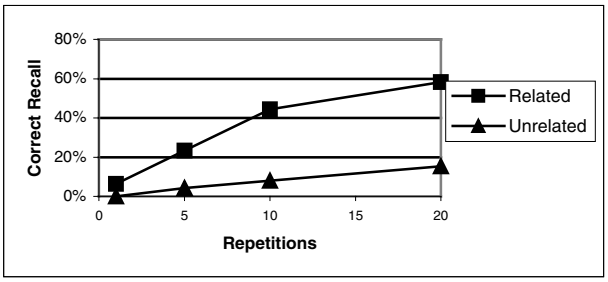


Figure 1: Percentages of correct recall for related and unrelated pairs in several learning repetition conditions. It is easier to form associations between related words.

These data demonstrated that semantic information is involved in forming episodic associations. For words that are semantically related, each learning repetition is more efficient. This facilitation is seen even if the subject's attention is not directed to the semantic level.

The aim of the present study was to develop a computational understanding of how associations are formed, including the influence of semantic factors on that process, as suggested by the above experiments.

Computational Model

Network Architecture

Our model is based on a Self-Organizing Semantic Map with lateral connections (Kohonen, 1995; Miikkulainen, 1992; Ritter & Kohonen, 1989). Semantic maps are 2-D networks that represent words by their nodes. The maps are formed by an unsupervised learning algorithm, such that words that are close in their meaning are represented by nearby nodes in the map. Hence, semantic relatedness is modeled by distance over the map. Semantic maps have been successfully used in various studies in which aspects of the semantic system were modeled, such as language acquisition (Li, 2000), semantic priming (Lowe, 1997), and semantic and episodic memory (Miikkulainen, 1992). Because self-organizing maps are based on biologically plausible Hebbian learning, and maps in general are common in many parts of the cortex (Knudsen, Lac & Esterly, 1987), self-organizing maps are most appealing as a biologically plausible analogue of classic semantic networks.

Based on a semantic map we added all-to-all unidirectional lateral connections to represent the potential associations between two words. The strength of each such connection is composed of semantic and episodic components:

$$Lat\left((i,j),(u,v)\right) = Sem\left((i,j),(u,v)\right) + Epis\left((i,j),(u,v)\right) \quad (1)$$

where $Lat\left((i,j),(u,v)\right)$ is the connection weight from node (u,v) to node (i,j). The semantic component represents the distance on the map and is given by the equation:

$$Sem\left((i,j),(u,v)\right) = 1/\left(1+e^{|w(i,j)-w(u,v)|}\right), \quad (2)$$

where $w(i,j)$ is the map's weights vector for neuron (i,j). Initially, the episodic part of all the lateral connections was set to zero. Hence, prior to any learning of associations, the lateral links only capture the topographic organization of the map, i.e. the semantic relatedness of words.

When a word is presented to the model, an activity bubble is generated surrounding the node that represents it. The activity wave then spreads according to synchronized recurrent dynamics. At each time step, the input to each neuron is the sum of the activities of all neurons in the previous time step, weighted by the lateral connections. Then, the neuron's activity is set according to a sigmoid function

$$S^{t}_{(i,j)} = \sigma\left(\sum_{(u,v)} Lat\left((i,j),(u,v)\right)\ S^{t-1}_{(u,v)}\right), \quad (3)$$

where

$$\sigma(x) = 1/\left(1+e^{-x}\right), \quad (4)$$

and $S^{t}_{(i,j)}$ is the activity of the neuron *(i, j)* at time *t*.

When two words are presented to the model (such as in the learning phase of Experiment 1 below) both activities spread independently over the map. The sum of the intersection of activation (the MIN of the two values) over all the map's neurons and over all time steps is calculated and added to the episodic component of the lateral connection between these two words. When the geometric distance between the two words is smaller (indicating stronger semantic relatedness), the resulting activity waves overlap more extensively, causing a greater amplification of the direct link between them. Thus, it is easier for the model to associate related words than unrelated words. Conceptually, this method is an abstraction of Hebbian learning of the associative links since the resulting connection strength depends on the intersection of both words' activation waves.

Input Representations

In order to organize the semantic map, we used numeric representations based on the lexical co-occurrence analysis in the Hyperspace Analogue to Language (HAL) model of Burgess and Lund (1997). These vectors have been shown to capture the semantics of words quite well (Burgess & Lund, 1997) and have been found successful in creating sensible Self-Organized Semantic Maps (Li, 2000). In the current simulations, HAL representations were based on the 3.8 million word CHILDES database, a corpus with particularly clearly defined word semantics.

The semantic map in our model consisted of 250 nouns organized on a 40 by 40 grid. We selected 48 nouns that formed 24 pairs of words, with the criterion that words in each pair belong to the same semantic category and thus are semantically related. The words were English translations of the 48 Hebrew words used in the behavioral experiment described above. In some cases, where a direct translation did not exist or the translation word did not appear in our set of HAL representations, a similar English word was selected. Another 202 nouns were selected randomly from the set of representations as "fillers" in the map, to create a richer semantic neighborhood in which the 48 words of interest could organize. See Figure 2 for the final semantic map that was used in the current simulations.

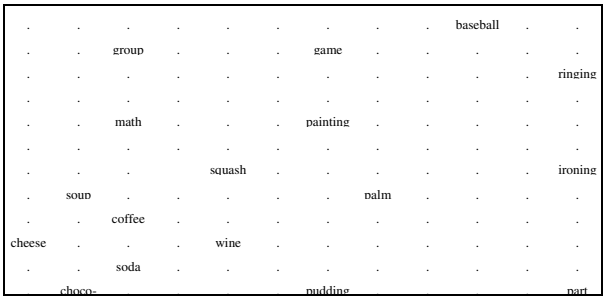


Figure 2: A section of the organized semantic map. Similar words are mapped to adjacent nodes.

Semantic Facilitation Simulation

The first simulation was aimed at replicating the empirical results of the second behavioral experiment that demonstrated semantic facilitation of associations' formation.

Experimental Setup

Out of the 24 semantically related pairs that were embedded in the semantic map, we selected 12 pairs that had numeric representations with a Euclidean distance shorter than the theoretical average (0.707) but larger then a threshold (0.5). Thus, these pairs were semantically related but not associated to each other prior to the experiment. In addition, we randomly re-matched the other 12 pairs such that 12 semantically unrelated pairs were formed as well.

Procedure

During the simulation of the learning phase of the experiment, in each trial the model was presented with two words with a certain time delay (i.e. SOA). Note that the absolute time scale of the network is arbitrary and can be adjusted to fit the data. Each of the 24 pairs (12 related and 12 unrelated) was presented once. Since during the learning phase, the episodic information does not affect the spreading activation process, the resulting association from multiple presentations was calculated simply by multiplying the result of a single presentation by the number of repetitions. The number of learning repetitions was varied from 1 to 30. During the simulation of the test phase, only one word was presented to the model. The resulting activation wave spread based on the same dynamics, except that in this phase, both the semantic and the episodic components of the lateral connections were taken into account. The activity continued to spread until the first neuron reached an activity threshold (0.98). The word represented by this node was then output as the result.

Results and Discussion

In Figure 3 the results that corresponds to the number of repetitions used in the behavioral experiment (1, 5, 10, 20) are shown. The percentages of correct recall demonstrated by the model for the related and unrelated pairs are shown for each repetition condition. As shown by the Figure, the model successfully replicates the results from the behavioral experiment. In the early stages, the learning rate of the related pairs is higher than the learning rate of the unrelated pairs. At about 10 repetitions, a ceiling effect reduces the learning rate of the related pairs, such that the advantage of these pairs over the unrelated pairs is abolished. In addition, the learning rate of the unrelated pairs is relatively constant. It is important to emphasize that non-linearity is introduced to the testing phase of the simulated experiment by the recurrent dynamics of the model. Hence, the linear way in which multiple repetitions were modeled does not dictate linearity in the output learning rate.

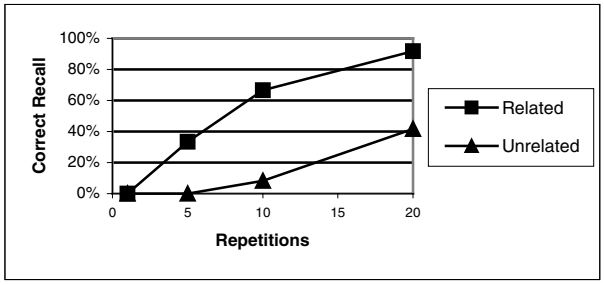


Figure 3: Percentages of correct recall demonstrated by the model, matching these obtained in the behavioral experiments (Figure 1).

Implicit Asymmetry Simulation

Associations between word pairs are directional. In free association questionnaires, for most pairs the subject would reply with word B after A with a different probability than the other way around (Koriat, 1981). In our model, this *explicit asymmetry* is achieved by the unidirectional lateral connections, which represent the association between two words in the map. However, our model demonstrates an additional asymmetry, which we call *implicit asymmetry*: it is sometimes easier to form an association between two words in one direction than in the opposite direction even before any episodic in-

formation is taken into account. The second simulation was aimed at quantifying this phenomenon.

Experimental Setup

First, we examined the density of the semantic neighborhoods of the words that were used in experiment 1. For each of the 48 words of interest we counted how many of the 250 total words in the model's semantic system were within a fixed 100-dimensional distance (0.4) according to their HAL representations. For each pair we then calculated the difference in the densities of the semantic neighborhoods of the two words and selected 3 related and 3 unrelated pairs with the greatest difference (in absolute values).

Procedure

We replicated the procedure of experiment 1 twice with the 6 pairs selected. First, the pairs were presented in the forward direction, from sparse to dense. Then, we repeated the entire procedure with the pairs presented in the opposite order (backward).

Results and Discussion

Figure 4 shows the percentages of correct recall as demonstrated by the model for pairs in the forward and backward direction in each repetition condition. Word pairs in the forward direction (sparser neighborhood → denser neighborhood) show advantage in correct recall as well as in learning speed throughout the first ten repetitions. The reason for this implicit asymmetry is the spreading of activation over a non-uniformly distributed high-dimensional space (elaborated in the General Discussion below). Although implicit asymmetry has not yet been observed experimentally, there is indirect evidence that suggests that such a process might indeed exist in the brain. Dagenbach, Horst and Carr (1990) found that it is easier to add a new word to semantic memory than to establish a link between two formerly unconnected words already in semantic memory. This result may apply to our prediction since we may assume that newly learned words were not yet well embedded into their semantic neighborhood and thus have a sparser semantic neighborhood than familiar words. In future work, we intend to test this prediction of the model with behavioral experiments.

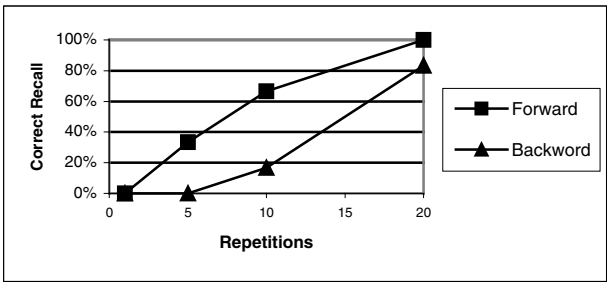


Figure 4: Percentages of correct recall demonstrated by the model for forward and backward pairs in several learning repetition conditions. The results show that learning is easier from sparse to dense neighborhoods.

General Discussion

As described in the "Behavioral Experiments" section, in a series of experiments we investigated semantic factors that affect the process of forming associations between words. The goal of this study was to present and evaluate a computational model that could account for these results and to produce further predictions regarding the process of associating words. Our model suggests that semantic relatedness between words, as well as episodic associations, could be implemented in a single structure using two distinct types of representations. On one hand, semantic relatedness is expressed as geometrical proximity in a high-dimensional (100D) feature space. On the other, episodic associations are represented by arbitrary "physical" connections between the units that represent the concepts. Both types of relations are implemented simultaneously in a semantic map with lateral connections artificial neural network.

As was demonstrated in human subjects, the model shows the facilitation semantic information has on learning new associations. This facilitation emerges in a natural, mechanistic manner, without involvement of top-down, intentional processes. It is achieved by implementing Hebbian link strengthening based on intersections of activation waves over a semantic map.

The asymmetric nature of relationships between words and more specifically of associations imposes difficulties for computational models that rely on geometric distances between high-dimensional numeric representations of words. Our model is also based on such high-dimensional vectors and the self-organization algorithm that establishes the semantic map is symmetric. Nonetheless, the model demonstrates two kinds of asymmetries. The first, explicit asymmetry is achieved by the unidirectional lateral connections that are implemented on top of the symmetric organization of the semantic map. These connections make it possible to have asymmetric associations between two words, based on the episodic experience of the two possible directions of the word pair. The second, implicit asymmetry, emerges from the non-uniform distribution of concepts in the high-dimensional space. This non-uniform distribution induces asymmetric distances in terms of spreading activation between two points in the semantic space that otherwise would have equal distance from one another in both directions.

Further empirical studies can be derived from this computational research. The model suggests that when an association is formed between two semantically distinct words, it can serve as a "pipeline" that enhances the spreading of activity from the semantic neighborhood of the first word to the semantic neighborhood of the second. Since this activity is, in turn, used to form other associations, we infer that the existence of an as-

sociation between words from distinct semantic neighborhoods (e.g. different categories) would facilitate forming other associations between unrelated pairs that belong to those semantic neighborhoods.

Another possible implication of this model is in testing one of the theories concerning *Schizophrenic Thought Disorder* (hereinafter STD) as suggested by Spitzer (1997). According to Spitzer's theory, the activation over the semantic network of STD patients spreads faster and further than that of normal subjects. This unfocused activation can explain experimental results in STD patients that show stronger semantic priming and indirect semantic priming. It may also explain the clinical STD phenomena of loose, oblique and derailed associations. By manipulating our model we can computationally test this theory. It is possible to vary the parameters of the functions that govern the spreading activation (equations 1-4) to make it less focused. By examining the resulting changes in the model's behavior, we may be able to gain insight regarding the processes that lead to this pathology.

Conclusion

We set out to study the process of creating new associations between words in human memory during incidental learning. Empirical results suggest that semantic information enhances the process of forming episodic associations. A model based on spreading activation on a laterally connected self-organizing map matches these results and leads to further insights into why such associations tend to be asymmetric. In future work, we plan to test some of the model's predictions, including implicit associations and processes of abnormal behavior.

Acknowledgements

We are grateful to Ping Li and Curt Burgess for providing the HAL vectors. This research has been supported in part by the National Science Foundation grant IIS-981147 to Dr. R. Miikkulainen, and by a German-Israeli Science Foundation grant #567 to Dr. S. Bentin.

References

Burgess, C. & Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, 12, 1-34.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82, 407-428.

Dagenbach, D., Horst, S., and Carr, T. H. (1990). Adding new information to semantic memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 16, 581-591.

Fischler, I. (1977). Semantic facilitation without association in a lexical decision task. *Memory & Cognition*, 5, 335-339.

Hinton, G. E. (1990). Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence*, 46, 47-75.

Hume, D. (1738/1962). A Treatise of Human Nature. London: J.M. Dent & Sons, Ltd.

Knudsen, E. I., Lac, S., & Esterly, S. D. (1987). Computational maps in the brain. *Annual Review of Neuroscience*, 10, 41-65.

Kohonen, T. (1995). *Self-Organizing Maps*. Springer.

Koriat, A. (1981). Semantic facilitation in lexical decision as a function of prime-target association. *Memory & Cognition*, 9, 587-598.

Li, P. (2000). The acquisition of tense-aspect morphology in a self-organizing feature map model. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp.304-309). Lawrence Erlbaum.

Lowe, W. (1997). Semantic representation and priming in a self-organizing lexicon. *Proceedings of the 4th Neural Computation and Psychology Workshop* (pp. 227-239). Springer-Verlag.

Masson, M. E. J. (1995). A Distributed Memory Model for Semantic Priming. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 21, 3-23.

McKoon, G. & Ratcliff, R. (1979). Priming in episodic and semantic memory. *Journal of verbal learning and verbal behavior*, 18, 463-480.

Meyer, D. E. & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words. *Journal of Experimental Psychology*, 90, 227-234.

Miikkulainen, R. (1992). Trace feature map. *Biological Cybernetics*, 66, 273-282.

Moss, H. E., Hare, M. L., Day, P., & Tyler, L. K. (1994). A distributed memory model of the associative boost in semantic priming. *Connection Science*, 6, 413-427.

Moss, H. E., Ostrin, R. K., Tyler, L. K., & Marslen-Wilson, W. D. (1995). Accessing different types of lexical semantic information: Evidence from Priming. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 21, 863-883.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner, & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition*. Hillsdale, NJ: Lawrence Erlbaum.

Plaut, D. C. (1995). Semantic and associative priming in a distributed attractor network. *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 37-42). Mahwah, NJ: Lawrence Erlbaum.

Ritter, H. & Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics*, 61, 241-254.

Shelton, J. R., & Martin, R. C. (1992). How semantic is semantic priming? *Journal of Experimental Psychology: Learning, Memory & Cognition*, 18, 1191-1210.

Silberman, Y. & Bentin, S. (submitted). *Semantic boost of forming episodic associations.*

Spitzer, M. (1997). A cognitive neuroscience view of schizophrenic thought disorder. *Schizophrenia Bulletin*, 23, 29-50.

Representation: Where Philosophy Goes When It Dies.

Peter Slezak (p.slezak@unsw.edu.au)
Program in Cognitive Science, University of New South Wales
Sydney NSW 2052 AUSTRALIA

Abstract

Robert Cummins (1996, p.1) has characterised the problem of mental representation as "*the* topic in the philosophy of mind for some time now". This remark is something of an understatement. The same topic was central to the famous controversy between Nicolas Malebranche and Antoine Arnauld in the Seventeenth Century and remained central to the entire philosophical tradition of "ideas" in the writings of Locke, Berkeley, Hume and Kant. I show that the recurrence of certain deep perplexities about the mind is a systematic and pervasive pattern, confirming Jerry Fodor's disparaging remark: "Cognitive science is where philosophy goes when it dies" (Fodor, 1994b, p. 110).

The Tripartite Schema

Recently Bechtel (1998, p. 299) states the essentials of a modern theory of representation: "There are ... three interrelated components in a representational story: what is represented, the representation, and the user of the representation".

Z: System Using Y → Y: Representation → X: Thing Represented

Among the problematic assumptions, Bechtel's diagram and discussion crucially fail to distinguish internal and external representations. Bechtel's conception in this regard is not idiosyncratic but almost universal in cognitive science (Newell 1986, p. 33). As we will note presently, in the case of pictorial images, the assimilation of internal and external representations tacitly encourages the illegitimate postulate of a user or external observer - the notorious homunculus. The same tacit assimilation of external and internal representations is at the heart of Searle's "refutation" of symbolic AI and also leads to the doctrine that we think "in" language (Carruthers 1996, Slezak forthcoming a). The assimilation just noted in Bechtel will also be seen in the seemingly unrelated problem of consciousness and the mind-body problem (Place 1956), *inter alia*.

The tripartite scheme appears obvious and innocuous enough but Bechtel's diagram (modified here) is a variant of the scheme which we see throughout the long history of the subject. Thus, for example, nothing could seem more remote from modern theories in cognitive science today than Malebranche's (1712/1997) seventeenth century doctrine of "the vision of all things in God". On the contrary, however, despite the theological trappings, it is instructive to recognize the profound affinity of Malebranche's views with those at the very forefront of theorising today in psychology and artificial intelligence: Malebranche's theory is just Bechtel's tripartite model (Nadler 1992), and the modern problem of representation is how to avoid the notorious difficulties clearly articulated by his critic Arnauld (1683/1990).

It is no accident that Gibson's 'ecological' approach involves a direct realism which has seen proposed as alternative to the representationalism of computational theories. This is merely one form in which the Malebranche-Arnauld debate is being rehearsed today. This celebrated debate is described by Nadler (1989) as a debate between an 'object theory' of ideas and an 'act theory', respectively. He explains

> ... the object theory of ideas involves a commitment to a representationalist or *indirect* realist theory of perception, such as Malebranche (and, on the traditional reading, Locke) put forth. An act theory of ideas, on the other hand, forms the core of Arnauld's perceptual direct realism. If ideas are representational mental acts [rather than entities], then they can put the mind in direct cognitive contact with the world - no intervening proxy, no *tertium quid*, gets in the way. (1989, p.6)

It is no accident that recent proponents of 'situated cognition' have been complaining of exactly the same indirect, mediated conception in computational theories of cognition. For example, J. Greeno (1989) unwittingly echoes Arnauld:

> I am persuaded ... we are connected directly with the environment, rather than connected indirectly through cognitive representations.
> ... An individual in ordinary circumstances is considered as interacting with the structures of situations directly, rather than constructing representations and interacting with the representations. (1989, p. 290)

Precursors: Pointless Exercise?

Despite the skepticism expressed by Gaukroger (1996), precursors of modern cognitive science provide an independent, extensive source of insight into contemporary issues and, conversely, are themselves elucidated in novel ways unavailable to traditional

scholarship (Yolton 1996, Slezak 1999, 2000). The possibility of two-way elucidation arises from the extraordinary persistence of the seemingly simple problem of saying "in some illuminating way, what it is for something in the mind to represent something" (Cummins 1996, p.1). From Yolton's (1984, 1996) statement of earlier concerns we see their parallel with the contemporary problem: Scholastics' notions of 'intelligible species' and Cartesian talk of ideas were striving for some way to explain the conformity or agreement between ideas and objects.

If Malebranche and Arnauld anticipated contemporary concerns about representation in cognitive science, then it is clear that the current theoretical problem has nothing to do with the theoretical framework of symbolic, computational approaches as universally assumed. Indeed, the recurrence of essentially the same dispute in widely varying contexts today suggests that the underlying problem does not arise essentially from the special features of any one of them. I suggest that we may discern the same underlying problem at the heart of notorious disputes such as 'The Imagery Debate', Searle's Chinese Room conundrum, the thinking-in-language debate, 'situated cognition' and a number of others which have been prominent and recalcitrant.

No Representations?

The 'Cognitive Revolution' was characterized by a re-discovery of the indispensability of internal representations following their repudiation by Skinnerian behaviourism. There is considerable irony in recent approaches which appear to reject internal representations once again (Brooks 1991, Freeman and Skarda 1990, Clark and Toribio 1994, Greeno 1989, van Gelder, 1998). Notwithstanding Eliasmith's (1996) claim, these views are not plausibly seen as a return to behaviourism, but they are symptoms of the profound difficulties posed by the phenomena. It is sobering to notice that Arnauld's critique of Malebranche exactly prefigures these recent attacks on representational theories. Arnauld's treatise *On True and False Ideas* is concerned to repudiate what he describes as "imaginary representations", saying "I can, I believe, show the falsity of the hypothesis of *representations*" (1683/1990, p.77) for "one must not make use of alleged *entities* of which we have no clear and distinct idea in order to explain the effects of nature, whether corporeal or spiritual" (1683/1990, p. 65).

Tables & Chairs: Bumping Into Things

Fodor (1985a) joked that philosophers are notorious for having been prey to absurd, eccentric worries such as the "fear that there is something fundamentally unsound about tables and chairs". Nevertheless, he optimistically opined that sometimes "mere" philosophical worries turn out to be *real* as in the case of the representational character of cognition. However, far from being a *contrast* with the traditional anxiety about tables and chairs, modern scientific disputes concerning representations appear to be *identical* with this notorious worry!

Thus, it is surely no accident that, reflecting upon Fodor's (1980) 'methodological solipsism', Jackendoff (1992, 161) asks facetiously "Why, if our understanding has no direct access to the real world, aren't we always bumping into things?" Though intending a mild parody, Jackendoff captures precisely the paradox charged against Locke and also Malebranche, who Nadler (1992, p. 7) says "is often portrayed ... as enclosing the mind in a "palace of ideas," forever cut off from any kind of cognitive or perceptual contact with the material world". Thus, Jackendoff's satire is evocative of Samuel Johnson's famous refutation of Berkeley's "ingenious sophistry" by kicking a stone. Of course, Berkeley's "sophistry" is just the worry about the reality of tables and chairs.

Imagery: The Pictorial Theory

The 'Imagery Debate' is perhaps the most remarkable modern duplication of seventeenth century controversies. In this re-enactment, among the *dramatis personae* Pylyshyn plays Arnauld against Kosslyn's Malebranche. Significantly, the central error identified by Arnauld of ascribing corporeal properties to mental ones exactly the one charged by Pylyshyn (1973, 1981) against Kosslyn and related to Cummins' (1996) point that internal representations do not function by being understood".

Kosslyn's (1994) pictorial account of imagery takes mental images to represent by virtue of a relation of resemblance to their objects and by virtue of actually *having* spatial properties which they represent. Furthermore, "depictive" representations in a "visual buffer" are taken to have the specific function of permitting a re-inspection of images by the higher visual apparatus. Not surprisingly, this "quasi-perceptual" model has been repeatedly charged with the error of importing an 'homunculus'. The charge is vigorously rejected on the grounds that "the theory is realized in a computer program" (Kosslyn, Pinker, Smith & Schwartz, 1979, p. 574), but undischarged homunculi can lurk in computational models just as easily as in traditional discursive theories (see Slezak 1992, 1994, 1995, 1999). Thus, Kosslyn, Sokolov and Chen (1989) offer a diagram of the visual imagery system which is a profusion of inter-connected boxes and arrows. The box labeled "visual buffer" contains another box labeled "attention window" which is left unexplained. This box is, in fact, the observer in the 'theater' which is the source of the traditional problem. The elaborate diagram is reducible to the same tripartite schema we have seen in Malebranche. Significantly, following Descartes, Arnauld explicitly pointed to the seductive error of taking pictures as an appropriate model of mental representation (Arnauld 1683/1990, p.

67) and he cites the *camera obscura* as an erroneous model for imagery. Thus, retinotopic maps on the visual cortex cited by Kosslyn (1994), p. 14) as vindicating the pictorial theory are pictures alright, but only for the *theorist.*

Descartes Déjà Vu.

A related unlikely indication of the relevance of early philosophy to current problems is seen in Edelman's (1998) work on perception. Despite its concern with the latest theories of perception, the central problem is stated in terms identical with that of the entire tradition of writers on 'ideas'. Edelman writes: "Advanced perceptual systems are faced with the problem of securing a principled (ideally, veridical) relationship between the world and its internal representation." Edelman's solution "is a call for the representation of similarity instead of representation by similarity". This might have been taken verbatim from Descartes's *Treatise of Man* or *Dioptrics* where he said "the problem is to know simply how [images] can enable the soul to have sensory perceptions of all the various qualities of the objects to which they correspond - not to know how they can resemble these objects" (Descartes 1985, 1, 165).

Mind-Body Problem

The pervasive error seen starkly in Kosslyn's TV screen metaphor reveals the link between the various problems in cognitive science and the traditional mind-body problem. In the classic statement of materialism, U.T. Place (1956) argued that the rejection of materialism is based on the qualitative features of subjective experience. Although these features have recently been supposed to constitute the "hard" problem of consciousness (Chalmers 1996), Place suggested that they are the source of the 'phenomenological fallacy'. This is "the mistake of supposing that when the subject describes his experience, how things look, sound, smell, taste, or feel to him, he is describing the literal properties of objects and events on a particular sort of internal cinema or television screen." Place's diagnosis has been revived and given prominence by Dennett (1991), however the fallacy would be more aptly named the 'Malebranchean Theater'.

Thinking In Language

Just as we seem to be looking at pictures when we imagine visually, so we appear to talk to ourselves when we think. Indeed, Carruthers (1996) who seeks to revive what he acknowledges to be an unfashionable doctrine explicitly bases his argument on such evidence of introspection. This is the evidence that we sometimes find ourselves in a silent monologue, talking to ourselves *sotto voce*, as it were.

However, in a neglected article, Ryle (1968) suggested that the very idea that we might think "in" language is unintelligible, and the undeniable experience of talking to ourselves cannot support any claim about the vehicles of thought. It is significant that Ryle mentions *en passant* among the equally problematical cases, that in which we claim to see things in our 'mind's eye' - taken to involve mental pictures of some kind. Ryle's comparison and his warning is unwittingly confirmed by Carruthers (1996, 1998) who explicitly invokes Kosslyn's pictorial account of imagery as support for his own analogous theory. In doing so, however, Carruthers only brings into relief the notorious difficulties of his own model which relies on representations - sentences of natural language - which are, like pictures, paradigmatically the kind requiring an external intelligent observer.

Connectionism & Cognitive Architecture

The crucial difference between symbolic and connectionist architectures is said to be the absence of *explicit* representations with constituent structure (Fodor & Pylyshyn 1988, Fodor & McLaughlin 1990). This issue may turn precisely on the same question we have seen. Specifically, the distinction between explicit and implicit representation appears to be based on a tacit appeal to the criterion of intelligibility or discernability to an external observer. D. Kirsh (1990, p. 340) points out that in connectionist systems, "it is becoming increasingly difficult … to track the trajectory of informational states these mechanisms generate. There is no doubt that we must find some method of tracking them; otherwise there is no reason to think of them as more than complex causal systems." Explicitness is characterised as a matter of "directly reading off" information from "visible structures" and the "immediate grasp" of information which is "directly available" or "immediately readable" (1990, p. 356), but the obvious question is: By whom? In an influential article Ramsey, Stich and Garon (1991) question traditional folk psychology because "in many connectionist networks it is not possible to localize propositional representation beyond the input layer" (1991, p. 209). Again we may ask, By whom? Significantly, Ramsey et al. imply that the difficulty of identifying representations in a neural net is "a real inconvenience to the connectionist model builder" (1991, p. 209). However, as Cummins' (1996, p.102) warns, "Internal representations are not exploited by being understood" by the programmer".

Symbols & Searle

Searle's (1980) Chinese Room conundrum appears to have an identical logical structure to those I have noted. In this case, a crucial equivocation on distinct meanings of 'meaning' has led to the postulation of symbols having meaning in an *observer-relative* sense in which a representation is necessarily apprehended and understood by someone. However, intelligibility *to the theorist* must be irrelevant to Searle's question of whether a

system has genuine, 'original' intentionality (see Slezak 1994, 1999). As before, rejecting this inappropriate criterion of meaning actually amounts to rejecting a certain conception of representations or, equivalently, rejecting the *agent* as homunculus in the system. Intentionality and the directness of cognition is achieved, following Arnauld, by eliminating a conception of symbols as intermediate objects to be apprehended as if they were external representations.

Searle's (1980) criterion for judging intentionality in his Chinese Room amounts to Carruthers' (1996) claim that the language of thought or 'mentalese' must be English, since the symbols are to be understood by a fully comprehending intelligent person. Accordingly, the much-discussed conundrum is best understood, not as a challenge to 'strong AI' as such, but as a *reductio ad absurdum* of symbols conceived as being intelligible to an observer. It is significant that this mistake is not Searle's alone, for it is implicit in the orthodox computational view of cognitive science which sees its origin in Frege-Russell formal symbols requiring an interpretation (Newell and Simon 1976, Newell, 1986). In AI, too, this conception has been explicitly embraced by Nilsson (1987, 1991) and embodied in the CYC program of Lenat and Feigenbaum which has been caricatured by Smith (1991) as the 'Electric Encyclopedia'. In these cases the question of meaning of mental representations is confused between whether representations are *intelligible* and whether they are *explainable*.

Searle's conundrum is evoked by Glanvill's response in 1661 to Descartes' coding theory of perception. "But how is it, and by what Art doth the soul read that such an image or stroke in matter … signifies such an object? Did we learn such an Alphabet in our Embryo-state?" (quoted in Yolton 1984, p. 28). Echoing Searle, Glanvill suggests that the "motions of the filaments of nerves" learn the quality of objects by analogy with the way in which a person learns to understand a language, for otherwise "the soul would be like an infant who hears sounds or sees lips move but has no understanding of what the sounds or movements signify, or like an illiterate person who sees letters but 'knows not what they mean' "(1984, p. 28).

Logicism & Observer Attribution

Within AI, an independent, though parallel, debate has been proceeding about the classical symbolic conception, the "logicist' view, according to which an abstract formal system gets its meaning from a model theory - that is, the intended interpretations of the designer (Newell and Simon 1976, Nilsson 1987, 1991). Woods (1987), Smith (1987), Rosenschein (1985) and others have argued that this classical logical view is fundamentally misguided in its conception of the way in which a system gets to relate to the external world by embracing the specific conception of "observer attribution" (see Hadley 1995). Birnbaum (1991, p. 62) says that AI mistakenly adopts a theory from logic which is quite inappropriate to capture what it means for a system to have beliefs.

Misrepresentation

The problem of misrepresentation has arisen for causal or co-variation theories of intentional content (Dretske 1996, Fodor 1994a) since these theories seem to be unable to capture the way a mismatch might arise between a representation and the world. If a mentalese token 'mouse' might be caused not only by mice but also by shrews, then the symbol must *ipso facto* mean 'shrew' and cannot be in error.

The puzzle might be accounted for by noting that it arises from tacitly adopting the stance of external interpreter: The very problem itself cannot be coherently formulated except in terms of judgements which are not part of the scientific, explanatory enterprise. The veridicality of representations is not a property which can play any role in the functioning of representations or the explanation of them. Like the picture on a jig-saw puzzle, the meaning of representations conceived as semantically evaluable in this way is for our own benefit and not intrinsic to the arrangements of inter-locking components.

The very concern with misrepresentation arises from tacitly adopting a questionable assumption endorsed by Davidson (1975) that having a belief requires also having the *concept* of belief, including the concept of error. However, it seems that animals might have beliefs even if they are unable to know that they have them and reflect on their truth value. A cat can surely be correct in thinking that a mouse is in a certain hole without having the *concepts* of belief and truth.

Argument From Illusion

The modern problem of misrepresentation is a variant of the classical 'argument from illusion' employed in support of Locke's 'ideas' and A.J. Ayer's (1940) sense-data as the immediate objects of perception. The parallel should not be surprising since an illusion in the relevant sense is precisely a misrepresentation. The traditional argument, just like Dretske's and Fodor's, turns on the possibility of a mismatch between mental representations and their referents in the external world.

Responding to Ayer, Austin (1962, p.61) remarked on the "curious" and "melancholy fact" that Ayer's position echoes that of Berkeley. Of course, this is the same melancholy fact that Fodor's "real" problems of representation are identical with the traditional concerns about the reality of tables and chairs. Questions of veridicality for ideas and sense-data arose from precisely the same assumptions as Fodor's - namely, the spurious possibility of a comparison between representations and the world. Twin Earth puzzles, too, seem to be an unnoticed variant on the problem of misrepresentation (Slezak, forthcoming b).

It should be less surprising that the classical arguments for 'ideas' should be akin to the modern case for representations when it is noticed that the 'argument from illusion' is effectively an 'argument from imagery'. The proverbial illusory pink elephant as the immediate object of perception is a visual image *par excellence*.

Illusion of the Intelligent Reader

We might expect a compelling kind of error to emerge in unrelated domains of theorising about the mind. Chomsky has drawn attention to the way in which traditional grammars produce an illusion of explanatory completeness while, in fact, they have "serious limitations so far as linguistic science is concerned" (Chomsky 1962, p 528). The success of the grammar depends on being "paired with an intelligent and comprehending reader". Here we see an entirely different version of the homunculus problem. Chomsky notes that in judging the adequacy of traditional grammars the unnoticed reliance on the user's linguistic ability is illegitimate because it is just what the theory is supposed to explain (Chomsky, 1962, p.528). Evidently, this is just the issue captured in Cummins' distinction, in a different context, between 'meaning' and 'meaningfor' (1996, p. 86) and is evidently the problem also for pictorial images or thinking in natural language.

Conclusion

Fodor (1968, p. vii) once remarked: "I think many philosophers secretly harbor the view that there is something deeply (ie. conceptually) wrong with psychology, but that a philosopher with a little training in the techniques of linguistic analysis and a free afternoon could straighten it out." Thirty years later, the suspicion of deep conceptual problems at the heart of philosophy and psychology is more clearly justified. By adopting a broader perspective we may see why the sorry fortunes of the two disciplines have been inextricably linked.

References

Arnauld, A. (1683/1990). *On True and False Ideas.* Trans. S. Gaukroger, Manchester University Press.

Austin, J.L. (1962). *Sense and Sensibilia*. Oxford University Press.

Ayer, A.J. (1940). *The Foundations of Empirical Knowledge.* Macmillan.

Bechtel, W. (1998). Representations and Cognitive Explanations. *Cognitive Science*, 22, 3, 295-318.

Birnbaum, L. (1991). Rigor Mortis: A Response to Nilsson's 'Logic and Artificial Intelligence'. *Artificial Intelligence*, 47, 57-77.

Brooks, R. (1991). Intelligence Without Representation. *Artificial Intelligence*, 47, 139-159.

Carruthers, P. (1996). *Language, Thought and Consciousness.* Cambridge University Press.

Carruthers, P. (1998). Thinking in Language? P. Carruthers and J. Boucher eds., *Language and Thought.* Cambridge University Press.

Chalmers, D. (1996). *The Conscious Mind.* Oxford University Press.

Chomsky, N. (1962). Explanatory Models in Linguistics. In E. Nagel. P. Suppes & A. Tarski, eds., *Logic, Methodology and Philosophy of Science*. Stanford University Press, 528-550.

Clark, A. & Toribio, J. (1994). Doing Without Representing? *Synthese*, 101, 3, 401-431.

Cummins, R. (1996). *Representations, Targets and Attitudes*. Bradford/MIT Press.

Davidson, D. (1975). Thought and Talk. In S. Guttenplan ed., *Mind and Language*. Clarendon Press.

Dennett, D.C. (1991). *Consciousness Explained.* Penguin.

Descartes, R. (1985). *The Philosophical Writings of Descartes*, in 2 Volumes. Translated by J. Cottingham, R. Stoothoff & D. Murdoch, Cambridge University Press.

Dretske, F. (1986). Misrepresentation. In S. Stich & T. Warfield eds., *Mental Representation.* Blackwell 1994.

Edelman, S. (1998). Representation is Representation of Similarities. *Behavioral and Brain Sciences,* 21, 449-498.

Eliasmith, C. (1996). The third contender. *Philosophical Psychology*, 9, 4, 441-463.

Fodor, J.A. (1968). *Psychological Explanation.* Random House.

Fodor, J.A. (1980). Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology. *Behavioral and Brain Science,* 3, 63-109.

Fodor, J.A. (1985a). Presentation. In B.H. Partee, S. Peters and R. Thomason eds., *Report of Workshop on Information and Representation*. NSF System Development Foundation, 106-117.

Fodor, J.A. (1985b). Fodor's Guide to Mental Representation. In *A Theory of Content and Other Essays*. MIT Press, 1990, 3-29.

Fodor, J.A. (1994a). *The Elm and the Expert,* MIT Press.

Fodor, J.A. (1994b). Concepts: A Potboiler. *Cognition*, 50, 95-113.

Fodor, J.A. & McLaughlin, B. (1990). Connectionism and the Problem of Systematicity. *Cognition*, 35, 183-204.

Fodor, J.A. & Pylyshyn, Z. (1988). Connectionism and Cognitive Architecture. *Cognition*, 28, 3-71.

Freeman, W.J. & Skarda, C.A. (1990). Representations: Who Needs Them? In J. L. McGaugh, N. Weinberger & G. Lynch eds. *Brain Organization and Memory Cells*. Oxford Univ Press.

Gaukroger, S. (1996). *Descartes: An Intellectual Biography*. Oxford University Press.

Greeno, J.G. (1989). Situations, Mental Models and Generative Knowledge. In D. Klahr and K. Kotovsky eds. *Complex Information Processing: The Impact of Herbert A. Simon.* Lawrence Erlbaum.

Hadley, R.F. (1995). The Explicit- Implicit Distinction. *Minds and Machines*, 5, 219-242.

Jackendoff, R. (1992). *Languages of the Mind*. Bradford/MIT Press.

Kirsh, D. (1990). When is Information Explcitly Represented? In P. Hanson, ed. *Information, Language and Cognition*. Univ of British Columbia Press.

Kosslyn, S.M. (1994). *Image and Brain: The Resolution of the Imagery Debate*. MIT Press.

Kosslyn, S.M., M.A. Sokolov & J.C. Chen 1989. The Lateralization of BRIAN. In D. Klahr & K. Kotovsky eds. *Complex Information Processing: The Impact of Herbert A. Simon*. Lawrence Erlbaum.

Kosslyn, S., Pinker, S., Smith, G. & Schwartz, S. (1979). On the demystification of mental imagery. *The Behavioral and Brain Sciences*, 2, 535-581.

Malebranche, N. (1712/1997). *The Search After Truth*. Trans. T.M. Lennon & P.J. Olscamp. Cambridge University Press.

Nadler, S. (1989). *Arnauld and the Cartesian Philosophy of Ideas.* Manchester University Press.

Nadler, S. (1992). *Malebranche and Ideas.* Oxford University Press.

Newell, A, (1986). The Symbol Level and the Knowledge Level. In Z. Pylyshyn and W. Demopoulos eds. *Meaning and Cognitive Structure.* Ablex.

Newell, A. & Simon, H.A. (1976). Computer Science as Empirical Inquiry. *Communications of the ACM,* 19, 113-126.

Nilsson, N.J. (1987). Commentary on McDermott. *Computational Intelligence,* 3, 202-203.

Nilsson, N.J. (1991). Logic and Artificial Intelligence. *Artificial Intelligence,* 47, 31-56.

Place, U.T. (1956). Is Consciousness A Brain Process?, In J. O'Connor ed., *Modern Materialism: Readings on Mind-Body Identity.* Harcourt Brace.

Pylyshyn, Z. (1973). What the Mind's Eye Tells the Mind's Brain. *Psychological Bulletin*, 80, 1, 1-24.

Pylyshyn, Z. (1981). The Imagery Debate. In N. Block, ed. *Imagery,* MIT Press.

Ramsey, W., Sitch, S. and Garon, J. (1991). Connectionism, Eliminativism and the Future of Folk Psychology. In W. Ramsey, S. Stich and D. Rumelhart, eds. *Philosophy and Connectionist Theory.* Lawrence Erlbaum.

Rosenschein, S.J. (1985). Formal Theories of Knowledge in AI and Robotics. *New Generation Computing*, 3, 345-357.

Ryle, G. (1968). A Puzzling Element in the Notion of Thinking. In P.F. Strawson ed., *Studies in the Philosophy of Thought and Action*. Oxford University Press, 7-23.

Searle, J. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences*, 3, 417-424.

Slezak, P. (1992). When Can Images Be Reinterpreted, *Proceedings of 14th Conference of the Society for Cognitive Science.* Lawrence Erlbaum, 124-129.

Slezak, P. (1994). Situated Cognition: Empirical Issue, Paradigm Shift or Conceptual Confusion. *Proceedings of 16th Conference of the Society for Cognitive Science.* Lawrence Erlbaum.

Slezak, P. (1995). The Philosophical Case Against Visual Imagery. In P. Slezak, T. Caelli and R. Clark eds. *Perspectives on Cognitive Science.* Ablex.

Slezak, P. (1999). Situated Cognition: Empirical Issue, Paradigm Shift or Conceptual Confusion? In J. Wiles & T. Dartnall eds. *Perspectives on Cognitive Science, Vol. 2*, Ablex.

Slezak P. (2000). Descartes' Startling Doctrine of the Reverse-Sign Relation. In S. Gaukroger, J. Schuster, J. Sutton eds. *Descartes' Natural Philosophy*. Routledge, 542-556.

Slezak, P. (forthcoming a). Thinking About Thinking, *Language and Communication.*

Slezak, P. (forthcoming b). Representing. In P. Staines, H. Clapin & P. Slezak eds. *Representation in Mind.* Greenwood.

Smith, B.C. (1987). The Correspondence Continuum. *CSLI Report 87-71*.

Smith, B.C. (1991). The Owl and the Electric Encyclopedia. *Artificial Intelligence,* 47, 251-288.

van Gelder, T. (1998). The Dynamical Hypothesis in Cognitive Science. *Behavioral and Brain Sciences*, 21, 615-665.

Woods, W.A. (1987). Don't Blame the Tools. *Computational Intelligence*, 3, 228-237.

Yolton, J.W. (1984). *Perceptual Acquaintance from Descartes to Reid.* University of Minnesota Press.

Yolton, J.W. (1996). *Perception and Reality:* Cornell University Press.

Effects of linguistic and perceptual information on categorization in young children

Vladimir M. Sloutsky (sloutsky.1@osu.edu)
Center for Cognitive Science & School of Teaching & Learning
Ohio State University
21 Page Hall, 1810 College Road
Columbus, OH 43210, USA

Anna V. Fisher (fisher.449@osu.edu)
Center for Cognitive Science & School of Teaching & Learning
Ohio State University
21 Page Hall, 1810 College Road
Columbus, OH 43210, USA

Abstract

This paper examines the process of categorization in young children, and tests predictions derived from a model of young children's similarity judgment. The model suggests that linguistic labels might have greater contribution to similarity judgment for younger children than do other attributes. It is argued that because categorization is based on similarity, the model predicting similarity judgment should also predict categorization. Predictions of the model were tested in the experiment where 4-6 year-olds were asked to perform a categorization task. Results of the experiment demonstrate that young children perform categorization in a similarity-based manner, and support both qualitative and quantitative predictions of the label-as-attribute model.

Introduction

The ability to group things together is an important component of human cognition: stimuli (i.e., objects, scenes, situations, or problems) rarely recur exactly, and, as a result, records of specific stimuli would be of little help. Therefore, the ability to form categories and store stimuli as members of these categories is a critical component of learning, memory, and thinking. Furthermore, it has been demonstrated that even infants are capable of forming categories (e.g., Balaban & Waxman, 1997; Quinn & Eimas, 1998; Mandler, 1997). It is less clear, however, how people form categories and include novel instances into a category. Several theories have emerged in an attempt to answer these questions (e.g., Lamberts & Shanks, 1997; Smith & Medin, 1984, for reviews).

The general question of how people form categories and add new instances to these categories consists of three more specific questions: (1) How do people decide whether or not a novel entity is a member of an existing category? (2) How do people form a category when presented with a large number of positive and negative instances of the category? And (3) how do people decide whether or not two novel entities are members of the same novel category? While much theoretical and empirical work on categorization has focused on the first two questions (see Lamberts & Shanks, 1997; Smith & Medin, 1984, for reviews), the third question has remained largely under-researched. At the same time, answers to this question are important for understanding of the "first step" in the process of categorization – forming a new category and including some novel entities as its members, while excluding others.

The current research attempts to examine the third question. As a first approximation, it appears plausible that, if no information about the entities is available, the entities would be grouped together on the basis of their perceptual similarity. If, in addition to perceptual information, there is also linguistic information (e.g., "*Look, here is an X*"), then there are at least two possibilities for grouping. If the label *X* is familiar, then the object denoted as *X* could be included into all categories that include *X* as its member. However, if label *X* is novel, it seems likely that categorization should be performed on the basis of similarity. In this case, a model predicting similarity judgment should also predict categorization. One such model, the label-as-attribute model suggests that young children consider linguistic labels as attributes of compared entities (Sloutsky & Lo, 1999). The model predicts that both perceptual and linguistic cues should contribute to comparison-based processes, such as similarity judgment. These predictions have been confirmed in a number of studies examining contribution of perceptual and linguistic factors to similarity judgment (Sloutsky & Lo, 1999) and inductive inference (Sloutsky & Lo, 2000; Sloutsky, Lo, & Fisher, in press). It was found that young children aggregate perceptual and linguistic cues when computing overall similarity among compared entities.

We can predict, therefore, categorization should be a function of similarity computed over perceptual and

linguistic cues. In what follows, we specify the model and its predictions, and present experiments designed to test predictions of the model.

The model is based on the product-rule model of similarity (Estes, 1994; Medin, 1975) that specifies similarity among non-labeled feature patterns. In the product-rule model, similarity is computed using Equation 1:

$$Sim(i, j) = S^{N-k} \qquad \mathbf{1}$$

where N denotes the total number of relevant attributes, k denotes the number of matches, and S ($0 \leq S \leq 1$) denotes values (weights) of a mismatch. For example, suppose that one is presented with two schematic faces A and B. Further suppose that these faces consist of four distinct features (i.e., the shape of the face, eyes, nose, and the size of ears), that they share two of these features (i.e., the shape of the face and eyes), and differ on the other two. Finally, suppose that $S = 0.5$, the value frequently derived empirically in past research (Estes, 1994). In this case, similarity between A and B would be equal to 0.25 (i.e., 0.5^2). Note that similarity between entities decreases very rapidly with a decrease in the number of mismatches, approximating the exponential decay function discussed elsewhere (Nosofsky, 1984). For example, if the faces share only one of the four features, their similarity would be equal to 0.125 (i.e., 0.5^3). On the other hand, if the faces share all four features, they would be identical, and their similarity would be equal to 1 (i.e., 0.5^0).

The label-as-attribute model suggest that linguistic labels might have greater contribution to similarity judgment for younger children than do other attributes, and there is evidence supporting this suggestion (see Sloutsky & Lo, 1999). Why would labels weigh more for younger children and what might be a mechanism underlying the greater weight of labels at earlier age? One possible explanation is that labels have larger weights because they are presented auditorily, and the auditory system matures earlier than the visual system. In particular, the auditory system starts functioning during the last trimester of gestation (Birnholz & Benaceraff, 1983; see also Jusczyk, 1998, for a review), whereas the visual system does not start functioning until after the birth. As a result, even though the neural bases of visual perception are fully developed at quite a young age (e.g., Aslin & Smith, 1988), auditory stimuli may still have a privileged processing status for younger children, thus resulting in larger weights of auditory stimuli (Napolitano, Sloutsky, & Boysen, 2001). In fact, it has been demonstrated that 15-months-olds grouped objects together when the objects shared an auditory input (either a label or a non-linguistic instrumental music input) if the input perfectly correlated with an infant's fixation of an object (Roberts, 1995; Roberts & Jacob, 1991, but see Balaban & Waxman, 1997).

According to the label-as-attribute model, similarity of labeled feature patterns could be calculated using Equation 2:

$$Sim(i, j) = S_{Label}^{1-L} S_{Vis.attr}^{N-k} \qquad \mathbf{2}$$

where N denotes the total number of visual attributes, k denotes the number of matches, $S_{vis.attr.}$ denotes values (attentional weights) of a mismatch on a visual attribute, S_{Label} denotes values of label mismatches, and L denotes a label match. When there is a label match, $L = 1$, and $S_{Label} = 1$; when there is a label mismatch, $L = 0$, and $S_{Label} < 1$. Note that S ($0 \leq S \leq 1$) denotes attentional weights of mismatches and the contribution of S is large if S is close to 0 and is small if S is close to 1. This is because the closer the value of S to 1, the smaller the contribution of a mismatch to the detection of difference, while the closer the value of S to 0, the greater its contribution to the detection of difference. When two entities are identical on all dimensions (i.e., there are no mismatches), their similarity should be equal to 1; otherwise, it is smaller than 1. Note that according to the model, when neither entity is labeled (i.e., $S_{Label} = 1$), similarity between entities is determined by the number of overlapping visual attributes, thus conforming to Equation 1. Labels are presented as a separate term in the equation because they are expected to have larger attentional weights than most visual attributes, an assumption that was borne out in previous research (Sloutsky & Lo, 1999). In the case that the weight of a label does not differ from that of other attributes, the label will become one of the attributes in the computation of similarity, and Equation 2 turns into Equation 1.

Finally, the model suggests that if the child is presented with a Target feature pattern (T) and Test feature patterns (A and B) and asked which of the Test patterns is more similar to the Target, the child's choices could be predicted using Equation 3:

$$P(B) = \frac{Sim(T, B)}{Sim(T, B) + Sim(T, A)} \qquad \mathbf{3}$$

In short, we argue that if categorization in young children is indeed similarity-based, then the same model that predicts similarity judgment in young children (e.g., Sloutsky & Lo, 1999) should be able to predict their categorization.

Simple derivations from Equation 3 allow us to predict categorization as a function of feature overlap. First, consider the case when entities are not labeled. Substituting *Sim (T,A)* and *Sim (T,B)* by their equivalents in Equation 1, we get Equation 4:

$$P(B) = S^x/(S^x + S^y) = S^x/[S^x(1 + S^{y-x})] = 1/(1 + S^y/S^x) \quad \mathbf{4}$$

For the labeled entities, derivations remain essentially the same, except for the S_{Label} parameter. The parameter equals to 1, if there is a label match, otherwise it equals to λ ($0 < \lambda < 1$). Therefore, in the case of labeled entities, the probability of selecting the item that shared the same label (say item B) could be derived according to Equation 5:

$$P(B) = S^x/(S^x + \lambda S^y) = S^x/[S^x(1 + \lambda S^{y-x})] = 1/(1 + \lambda S^y/S^x) \quad \mathbf{5}$$

In short, in the no-label condition, the probability of categorizing Test B and the Target together should be a function of the ratio of S^y/S^x (i.e., the ratio of similarity of Test A and Test B to the Target), whereas in the label condition such categorization should be a joint function of S^y/S^x and λ (i.e., attentional weight of label). Because we can estimate λ from our prior research, we can use Equations 4 and 5 for estimating specific probabilities of categorization. One important (and testable) consequence of this proposal is that because linguistic labels contribute to similarity in a quantitative manner rather than in a qualitative "all-or-nothing" manner, they should also make a quantitative contribution to categorization.

When stimuli consist of a small number of easily distinguishable and countable features (e.g., schematic faces or dot patterns), *N* and *K* (Equations 1 and 2) and subsequently *X* and *Y* (Equations 4 and 5) could be computed directly. However, if stimuli are perceptually rich, the task of determining *N*, *K*, and subsequently *X* and *Y* is complicated, if not impossible. One possible solution to this problem is to conduct a calibration study estimating similarity of each of the Test stimuli to the Target. Because similarity of each of the two Test stimuli to the Target is equivalent to S^y and S^x, ratios of similarity (i.e., S^y/S^x) could be easily computed, and therefore could be used to test the model.

The overall experimental idea was as follows. Participants were presented with triads of stimuli, with each triad consisting of Test stimuli A and B and Target T. In order to use perceptually rich stimuli and to quantify perceptual similarity, the stimuli were selected from sequences of images, in which one animal was "morphed" into another in a fixed number of steps. An example of a morphed sequence is presented in Figure 1. Multiple triads were formed from these sequences. These triads were subjected to a preliminary "calibration" study, in which participants were asked to estimate similarity of each of the Test stimuli to the Target. Those triads that gave the ratios of .5/.5, .4/.6, .3/.7, and .1/.9 were selected for the major study. An example of a .3/.7 triad is presented in Figure 2.

In addition to the quantitative predictions of Equations 4 and 5, we can formulate two qualitative predictions:

(1) When entities are not labeled, the probability of categorizing of Test stimulus together with the Target is a function of the ratio of perceptual similarity of this Test stimulus and the competing test stimulus to the Target.

(2) When entities are labeled, linguistic labels should affect categorization in a quantitative manner rather than in a qualitative "all-or-nothing" manner. Categorization should be a function of two variables – the weight of linguistic label and the similarity ratios – and not of linguistic labels alone.

Method

Participants

Participants were 37 preschool children recruited from daycare centers located in middle class suburbs of Columbus, Ohio (19 girls and 18 boys, $M = 5.4$ years; $SD = 0.82$ years).

Materials and Design

The experiment had a mixed design with a labeling condition (label vs. no-label) as a between-subject variable and similarity ratio as a within-subject variable. At both levels of the labeling condition participants were presented with the same triads of animal faces, one of which was a Target and two of which were Test stimuli. The only difference between the levels of the labeling condition was that in the label condition all stimuli were labeled, whereas in the no-label condition these stimuli were not labeled.

Materials consisted of triads of 4" × 4" pictures of animal faces selected to represent four levels of the stimulus pattern condition. Selection was made on the basis of results obtained in the calibration study. Each triad of pictures included a Target and two Test stimuli. The Target was located at the center above the Test stimuli.

These stimuli were selected by conducting a calibration experiment, in which 19 4–5 year-old children were presented with triads of pictures of animals (similar to those in Figure2) and asked which of the Test stimuli was more similar to the Target. 16 triads were selected on the basis of this calibration, representing four similarity ratios of (e.g., *Sim* (A, Target)/ *Sim* (B, Target): (1) .5/.5 = 1, (2) .4/.6 = 1.5, (3) .3/.7 = 2.33, and (4) .1/.9 = 9. Each of the four ratios included 4 triads. These four levels of perceptual similarity were included in the design.

Figure 1. Examples of 5 steps in a 20-step morphing sequence.

Step 1 **Step 6** **Step 10** **Step 14** **Step 18**

Figure 2. Example of an experimental triad

Target

Sim to the Target = .7

Sim to the Target = .3

Procedure

Triads of pictures were presented to each participant on a computer screen. A female researcher interviewed each child individually in a quiet room in their schools. Before the experimental task participants were introduced to two warm-up trials. Questions asked during the warm-up trials were identical to the questions asked during the experimental trials. No feedback was given to the participants on their performance on the warm-up or experimental trials, and no participant was eliminated from the study on basis of his/her performance in the warm-up. The sole purpose of the warm-up was to illustrate to children the nature of the task they were to perform.

Experimental trials were identical to warm-up trials. In the label condition participants were first introduced to the labels for the Target and Test stimuli and asked to repeat them. All the labels used were two-syllable artificial count nouns (e.g. a Bala, a Guga). No labels were introduced in the no-label condition. Then, children were asked whether the Target was the same kind of animal as Test 1 or Test 2. Positions of two Test stimuli were randomized across trials. In both conditions participants had 16 experimental trials (four trials each of the four within-subject stimulus patterns). The order of trials was randomized for each participant.

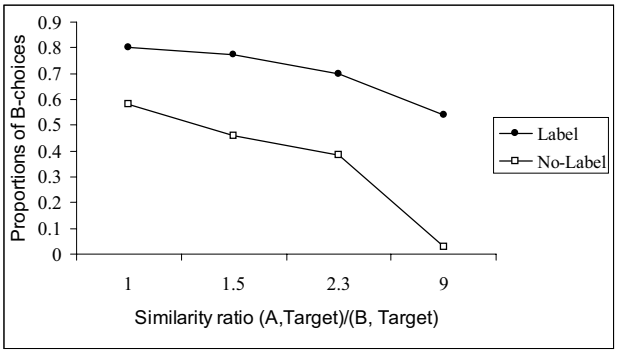
The important part of the instruction for preschool participants read: *Now we are going to play a game about animals from other planets. I am going to show you pictures of those alien animals, tell you their names, ask you to remember their names, and repeat them to me. Then I will ask you one question about those animals. Are you ready to start? I will show you something like this* (a warm-up triad was introduced at this point). *Look at them: this is a Guga* (points to the Target). *This is a Bala* (points to Test A). *This is a Guga* (points to Test B). *Could you please repeat their names? Do you think that this Guga* (points to the Target) *is the same kind as this Bala* (points to Test A) *or this Guga* (points to Test B)?

Note that in the no-label condition all stimuli were referred to as "this one." The order of introduction of the Test stimuli and their location relative to the target were randomized.

Results and Discussion

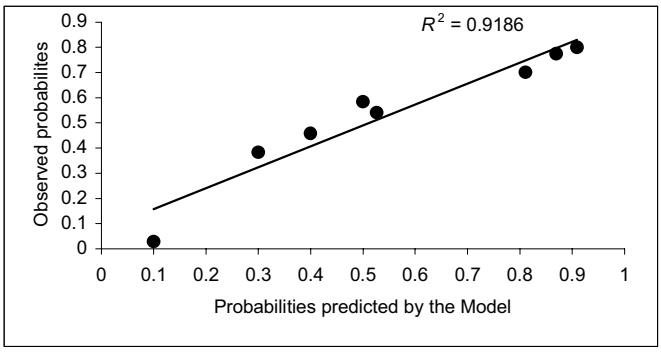
Proportions of B-choices by levels of the similarity ratio and labeling condition are presented in Figure 3 (recall that in the Label condition, Test B always shared the label with the Target). These proportions were averaged across the four trials for each level of the ratio and then averaged across subjects. Proportions averaged across trials were then subjected to a two-way (Labeling condition by Similarity ratio) mixed ANOVA with levels of similarity ratio as a repeated measure. The analyses indicated a significant main effect of labeling ($M_{\text{Label}} = .70 > M_{\text{No-Label}} = .36$), $F(1,35) = 30$, $MSE = 0.15$, $p < .0001$, and a significant main effect of the similarity ratio, $F(3,105) = 22.9$, $MSE = 0.06$, $p < .0001$, with no significant interaction. Planned comparisons of the levels of similarity ratio pointed to the following direction P(1) > P(1.5) = P(2.33) > P(9), all ts > 2, ps < .05. These results support the qualitative predictions, indicating that (a) when entities are not labeled, categorization is a function of perceptual similarity; (b) when entities are labeled, categorization is a function of similarity computed over perceptual and linguistic cues, and (c) labels contribute quantitatively to similarity among entities.

Figure 3. Proportions of B-choices by similarity ratio and labeling condition.



Quantitative predictions of the model are presented in Figure 4, where predicted probabilities of B-choices are plotted against observed probabilities. For the no-label condition, these probabilities were derived from Equation 4, whereas for the label condition they were derived from Equation 5 ($\lambda = .1$ was estimated from previous Sloutsky & Lo's data sets). Results indicate a good fit between predicted and observed probabilities ($r = .95$) with approximately 92% of variance explained by the model. These results indicating that similarity predicts much of categorization in young children support the hypothesis that, at least when labels are novel, categorization in young children is a function of similarity.

Figure 4. Overall fit of the model.



Several issues, however, would require further research. In particular, it remains unclear whether or not adults exhibit the same pattern of categorization as children. On the one hand, if entities are novel, it seems likely that adults should also use similarity as a basis of their categorization. On the other hand, there is evidence (Sloutsky, Lo, & Fisher, in press; Yamauchi & Markman, 2000) that adults are more likely than children to consider linguistic labels as category markers. There is also evidence that under different conditions adults may either rely on similarity for categorization (Smith & Sloman, 1994), or ignore it (Rips, 1989).

In short, the reported results support both quantitative and qualitative predictions of the model of label-as-attribute. As predicted, categorization appeared to be a function of two variables – the weight of linguistic label and the similarity ratios – and not of linguistic labels alone. These results also support the contention of the model that for young children linguistic labels are distinct attributes of entities. High correlations between the probabilities predicted by the model of similarity and the observed categorization frequencies support the hypothesis that categorization in young children is a similarity-based process.

Acknowledgments

This research has been supported by grants from the James S. McDonnell Foundation and the National Science Foundation to the first author.

References

Aslin, R., & Smith, L. (1988). Perceptual Development. *Annual Review of Psychology*, 39, 435-474.

Balaban, M. T., & Waxman, S. R. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child Psychology*, *64*, 3-26.

Birnholz, J. C., & Benaceraff, B. B. (1983). The development of human fetal hearing. *Science*, *222*, 516-518.

Estes, W. K. (1994). *Classification and cognition*. New York: Oxford University Press.

Jusczyk, P. W. (1998). *The discovery of spoken language*. Cambridge, MA: MIT Press.

Lamberts, K., & Shanks, D. R. (Eds), (1997). *Knowledge, concepts and categories*. Cambridge, MA: MIT Press.

Mandler, J. M. (1997). Development of categorisation: Perceptual and conceptual categories. In G. Bremner, A. Slater, & G. Butterworth (Eds.), *Infant development: Recent advances* (pp. 163-189). Hove, England UK: Psychology Press.

Medin, D. (1975). A theory of context in discrimination learning. In G. Bower (Ed.), *The psychology of learning and motivation* (pp. 263-314), Vol. 9. New York: Academic Press.

Medin, D. L., & Smith, E. E. (1984). Concepts and concept formation. *Annual Review of Psychology*, *35*, 113-138.

Napolitano, A., Sloutsky, V. M., & Boysen, S. (2001). *Proceedings of the XXIII Annual Conference of the Cognitive Science Society*.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, memory, & Cognition*, *10*, 104–114.

Quinn, P., & Eimas, P. (1998). Evidence for global categorical representation of humans by young infants. *Journal of Experimental Child Psychology*, *69*, 151-174.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (21-59). New York: Cambridge University Press.

Roberts, K. (1995). Categorical responding in 15-month-olds: Influence of the noun-category bias and the covariation between visual fixation and auditory input. *Cognitive Development*, *10*, 21-41.

Roberts, K., & Jacob, M. (1991). Linguistic vs. attentional influences on nonlinguistic categorization in 15-months-old infants. *Cognitive Development*, *6*, 355-375.

Sloutsky, V. M., & Lo, Y.-F. (1999). How much does a shared name make things similar? Part 1: Linguistic labels and the development of similarity judgment. *Developmental Psychology*, *6*, 1478-1492.

Sloutsky, V. M., & Lo, Y.-F. (2000). Linguistic labels and the development of inductive Inference. *Proceedings of the XXII Annual Conference of the Cognitive Science Society* (pp. 469-474). Mahwah, NJ: Erlbaum.

Sloutsky, V. M., Lo, Y.-F., & Fisher, A. V. (In Press). How much does a shared name make things similar? Linguistic labels and the development of inductive inference. *Child Development*.

Smith, E. E.& Sloman, S. A. (1994). Similarity- versus rule-based categorization. *Memory & Cognition*, 22, 377-386.

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *26*, 776-795.

The Interaction of Explicit and Implicit Learning: An Integrated Model

Paul Slusarz (kmicic@ideaworks.com)
Ron Sun (rsun@cecs.missouri.edu)
Department of CECS
University of Missouri-Columbia
Columbia, MO 65211, USA

Abstract

This paper explicates the interaction between the implicit and explicit learning processes in skill acquisition, contrary to the common tendency in the literature of studying each type of learning in isolation. It highlights the interaction between the two types of processes and its various effects on learning, including the synergy effect. This work advocates an integrated model of skill learning that takes into account both implicit and explicit processes; moreover, it embodies a bottom-up approach (first learning implicit knowledge and then explicit knowledge on its basis) towards skill learning. The paper shows that this approach accounts for various effects in the process control task data, in addition to accounting for other data reported elsewhere.

Introduction

The role of implicit learning in skill acquisition and the distinction between implicit and explicit learning have been widely recognized in recent years (see, e.g., Reber 1989, Stanley et al 1989, Willingham et al 1989, Proctor and Dutta 1995, Anderson 1993), Although implicit learning has been actively investigated, complex and multifaceted interaction between the implicit and the explicit and the importance of this interaction have not been universally recognized; to a large extent, such interaction has been downplayed or ignored, with only a few notable exceptions. Research has been focused on showing the *lack* of explicit learning in various learning settings (see especially Lewicki et al 1987) and on the controversies stemming from such claims. Similar oversight is also evident in computational simulation models of implicit learning (with few exceptions such as Cleeremans 1994).

Despite the lack of studies of interaction, it has been gaining recognition that it is difficult, if not impossible, to find a situation in which only one type of learning is engaged (Reber 1989, Seger 1994, but see Lewicki et al 1987). Our review of existing data (see Sun et al 2001) has indicated that, while one can manipulate conditions to emphasize one or the other type, in most situations, both types of learning are involved, with varying amounts of contributions from each (see, e.g., Sun et al 2001; see also Stanley et al 1989, Willingham et al 1989).

Likewise, in the development of cognitive architectures (e.g., Rosenbloom et al 1993, Anderson 1993), the distinction between procedural and declarative knowledge has been proposed for a long time, and advocated or adopted by many in the field (see especially Anderson 1993). The distinction maps roughly onto the distinction between the explicit and implicit knowledge, because procedural knowledge is generally inaccessible while declarative knowledge is generally accessible and thus explicit. However, in work on cognitive architectures, focus has been almost exclusively on "top-down" models (that is, learning first explicit knowledge and then implicit knowledge on the basis of the former), the bottom-up direction (that is, learning first implicit knowledge and then explicit knowledge, or learning both in parallel) has been largely ignored, paralleling and reflecting the related neglect of the interaction of explicit and implicit processes in the skill learning literature. However, there are a few scattered pieces of work that did demonstrate the parallel development of the two types of knowledge or the extraction of explicit knowledge from implicit knowledge (e.g, Rabinowitz and Goldberg 1995, Willingham et al 1989, Stanley et al 1989), contrary to usual top-down approaches in developing cognitive architectures.

Many issues arise with regard to the interaction between implicit and explicit processes: (1) How can we best capture implicit and explicit processes computationally? (2) How do the two types of knowledge develop along side each other and influence each other's development? (3) How is bottom-up learning possible and how can it be realized computationally? (4) How do the two types of knowledge interact during skilled performance and what is the impact of that interaction on performance? For example, the synergy of the two may be produced, as in Sun et al (2001). In this paper, we will focus on the interaction and the synergy resulting from the interaction.

A Model

Let us look into a model that incorporates both implicit and explicit processes.

Representation. The inaccessible nature of implicit knowledge may be captured by subsymbolic distributed representations provided by a backpropagation network (Rumelhart et al 1986). This is because representational units in a distributed representation are capable of accomplishing tasks but are subsymbolic and generally not

individually meaningful (see Rumelhart et al 1986, Sun 1995); that is, they generally do not have an associated semantic label. This characteristic of distributed representation accords well with the inaccessibility of implicit knowledge. [1] In contrast, explicit knowledge may be captured in computational modeling by a symbolic or localist representations (Clark and Karmiloff-Smith 1993), in which each unit is easily interpretable and has a clear conceptual meaning, i.e., a semantic label. This characteristic captures the property of explicit knowledge being accessible and manipulable (Smolensky 1988, Sun 1995). This radical difference in the representations of the two types of knowledge leads to a two-level model CLARION (which stands for *Connectionist Learning with Adaptive Rule Induction ON-line*; proposed in Sun 1997), whereby each level using one kind of representation captures one corresponding type of process (either implicit or explicit). [2]

Learning. The learning of implicit action-centered knowledge at the bottom level can be done in a variety of ways consistent with the nature of distributed representations. In the learning settings where correct input/output mappings are available, straight backpropagation (a supervised learning algorithm) can be used for the network (Rumelhart et al 1986). Such supervised learning procedures require the a priori determination of a uniquely correct output for each input. In the learning settings where there is no input/output mapping externally provided, reinforcement learning can be used (Watkins 1989), especially Q-learning (Watkins 1989) implemented using backpropagation networks. Such learning methods are cognitively justified: e.g., Shanks (1993) showed that human instrumental conditioning (a simple type of skill learning) was best captured by associative models (i.e., neural networks), when compared with a variety of rule-based models. Cleeremans (1997) argued that implicit learning could not be captured by symbolic models.

Specifically, $Q(x,a)$ is the "quality value" of action a in state x, output from a backpropagation network. Actions can be selected based on Q values, for example, using the Boltzmann distribution (Watkins 1989).

We learn the Q value function as follows:

$$\Delta Q(x,a) = \alpha(r \quad \gamma \max_b Q(y,b) - Q(x,a)) = \alpha(r - Q(x,a))$$

where x is the current state, a is one of the action. r is the immediate reward, and $\gamma \max_b Q(y,b)$ is set to zero for the process control task we tackle in this paper, because we rely on immediate reward in this particular task (details below). $\Delta Q(x,a)$ provides the error signal needed by the backpropagation algorithm and then backpropagation takes place. That is, learning is based on minimizing the following error at each step:

$$err_i = \begin{cases} r - Q(x,a) & \text{if } a_i = a \\ 0 & \text{otherwise} \end{cases}$$

where i is the index for an output node representing the action a_i. Based on the above error measure, the backpropagation algorithm is applied to adjust internal weights (which are randomly initialized before training).

The action-centered explicit knowledge at the top level can also be learned in a variety of ways in accordance with the localist representations used. Because of the representational characteristics, one-shot learning based on hypothesis testing (Nosofsky et al 1994, Sun 1997) is needed. With such learning, individuals explore the world, and dynamically acquire representations and modify them as needed, reflecting the dynamic (on-going) nature of skill learning (Sun 1997, Sun et al 2001). The implicit knowledge already acquired in the bottom level can be utilized in learning explicit knowledge (through *bottom-up* learning; Sun et al 2001).

Initially, we hypothesize rules of a certain form to be tested (Dienes and Fahey 1995, Nosofsky et al 1994). When a measure of a rule (the IG measure) falls below the deletion threshold, we delete the rule. Whenever all the rules of a certain form are deleted, a new set of rules of a different form are hypothesized, and the cycle repeats itself. In hypothesizing rules, we progress from the simplest rule form to the most complex, in the order as shown in Figure 1, in accordance with those numerical relations used in human experiments (Berry and Broadbent 1988, Stanley et al 1989). (Other rule forms can be easily added to the hypothesis testing process. Since rules are tested in a parallel fashion, adding more rules will not drastically change the working of the model.)

The IG measure of a rule is calculated (in this process control task) based on the immediate reward at every step when the rule is applied. The inequality, $r > threshold$, determines the positivity/negativity of a step and of the rule matching this step. [3] Then, PM (positive match) and NM (negative match) counts of the matching rules are updated. IG is then calculated based on PM and NM:

$$IG(C) = log_2 \frac{PM(C) \quad c_1}{PM(C) \quad NM(C) \quad c_2}$$

where C is the current rule and c_1 and c_2 (where $2 \quad c_1 = c_2$) are Laplace estimation parameters. Thus, IG essentially measures the positive match ratio of a rule.

Simulation of human skill learning data

Simulation Focus. A number of well known skill learning tasks that involve both implicit and explicit processes were chosen to be simulated that span the spectrum ranging from simple reactive skills to more complex cognitive skills. The tasks include serial reaction time tasks,

[1] However, it is generally not the case that distributed representations are not accessible at all but they are definitely less accessible, not as direct and immediate as localist representations. Distributed representations may be accessed through indirect, transformational processes.

[2] Sun (1995, 1997), and Smolensky (1988) contain more theoretical arguments for such two-level models (which we will not get into here).

[3] In the process control task, $r = 1$ if *process-outcome* = *target+/-1* and $r = 0$ otherwise, and $threshold = 0.9$.

$P = aW$	b
$P = aW_1$	b
$P = aW$	cP_1
$P = aW_1$	bP_2

Figure 1: The order of rules to be tested. $a = 1, 2$, $b = -1, -2, 0, 1, 2$, $c = -1, -2, 1, 2$, P is the desired system output level (the goal), W is the current input to the system (to be determined), W_1 is the previous input to the system, P_1 is the previous system output level (under W_1), and P_2 is the system output level at the time step before P_1.

process control tasks, the Tower of Hanoi task, and the minefield navigation task.

We focus on simulating process control tasks in this paper. We are especially interested in capturing the interaction of the two levels in the human data, whereby the respective contributions of the two levels are discernible through various experimental manipulations of learning settings that place differential emphases on the two levels. These data can be captured using the two-level interactive perspective.

We aim to capture (1) the verbalization effect, (2) the explicit (how-to) instruction effect, and (3) the explicit search effect. Through the simulations, it will be shown that the division of labor between, and the interaction of, the two levels is important.

To capture each individual manipulation, we do the following: (1) The explicit (how-to) instructions condition is modeled using the explicit encoding of the given knowledge at the top level (prior to training). (2) The verbalization condition (in which subjects are asked to explain their thinking while or between performing the task) is captured in simulation through changes in parameter values that encourage more top-level activities, consistent with the existing understanding of the effect of verbalization (that is, subjects become more explicit; Stanley et al 1989, Sun et al 1998). (3) The explicit search condition (in which subjects are told to perform an explicit search for regularities in stimuli) is captured through relying more on the (increased) top-level rule learning, in correspondence with what we normally observe in subjects under the kind of instruction. (4) Many of these afore-enumerated manipulations lead to what we called the synergy effect between implicit and explicit processes: that is, the co-existence and interaction of the two types of processes leads to better performance than either one alone (Sun et al 2001). By modeling these manipulations, we at the same time capture the synergy effect as well.

General Model Setup. Many parameters in the model were set uniformly as follows: Network weights were randomly initialized between -0.01 and 0.01. Percentage combination of the two levels (through a weighted sum) is used: that is, if the top level indicates that action a has an activation value l_a (which should be 0 or 1 as rules are binary) and the bottom level indicates that a has an activation value q_a (the Q-value), then the final outcome is $v_a = w_1 \quad l_a \quad w_2 \quad q_a$. The combination weights of the two levels were set at $w_1 = 0.2$ and $w_2 = 0.8$. Stochastic decision making with the Boltzmann distribution (based on the weighted sums) is then performed to select an action out of all the possible actions. The Boltzmann distribution is as follows:

$$p(a\ x) = \frac{e^{v_a\ \alpha}}{\sum_i e^{v_{a_i}\ \alpha}}$$

Here α controls the degree of randomness (temperature) of the decision-making process. It was set at 0.01. (This method is also known as Luce's choice axiom.) Other parameters include numbers of input, output, and hidden units, the external reward, the rule deletion threshold, the backpropagation learning rate, and the momentum. Most of these parameters were not free parameters, because they were set in an a priori manner (based on our previous work), and not varied to match the human data.

For modeling each of these manipulations, usually only one or a few parameter values are changed. These parameters are changed as follows. To capture the verbalization effect, we raise the rule deletion threshold at the top level. The hypothesis is that, as explained earlier, verbalization tends to increase top-level activities, especially rule learning activities. To capture the explicit search effect, we increase the weighting of the top level in addition to raising the rule deletion threshold. The hypothesis is that explicit search instructions tend to increase the reliance on top-level rule learning. To capture the explicit instruction effect, we simply wire up explicit a priori knowledge at the top level.

Simulating Stanley et al (1989)

The task. Two versions of the process control task were used in Stanley et al (1989). In the "person" version, subjects were to interact with a computer simulated "person" whose behavior ranged from "very rude" to "loving" (over a total of 12 levels) and the task was to maintain the behavior at "very friendly" by controlling his/her own behavior (which could also range over the 12 levels, from "very rude" to "loving"). In the sugar production factory version, subjects were to interact with a simulated factory to maintain a particular production level (out of a total of 12 possible production levels), through adjusting the size of the workforce (which has 12 levels). In either case, the behavior of the simulated system was determined by $P = 2 \quad W - P_1 \quad N$, where P was the current system output, P_1 was the previous system output, W was the subjects' input to the system, and N was noise. Noise (N) was added to the output of the system, so that there was a chance of being up or down one level (a 33% chance respectively).

There were four groups of subjects. The control group was not given any explicit how-to instruction and not asked to verbalize. The "original" group was required to verbalize: Subjects were asked to verbalize after each block of 10 trials. Other groups of subjects were

human data

	sugar task	person task
control	1.97	2.85
original	2.57	3.75
memory training	4.63	5.33
simple rule	4.00	5.91

Figure 2: The human data for the process control task from Stanley et al (1989).

model data

	sugar task	person task
control	2.276	2.610
original	2.952	4.187
memory training	4.089	5.425
simple rule	4.073	5.073

Figure 3: The model data for the task of Stanley et al (1989).

given explicit instructions in various forms, for example, "memory training", in which a series of 12 correct input/output pairs was presented to subjects, or "simple rules", in which a simple heuristic rule ("always select the response level half way between the current production level and the target level") was given to subjects. The numbers of subjects varied across groups. 12 to 31 subjects were tested in each group. All the subjects were trained for 200 trials (20 blocks of 10 trials).

The data. The exact target value plus/minus one level (that is, "friendly", "very friendly", or "affectionate") was considered on target. The mean scores (numbers of on-target responses) per trial block for all groups were calculated. Analysis showed the verbalization effect: The score for the original group was significantly higher than the control group ($F(1,73) = 5.20, p < 0.05$). Analysis also showed the explicit instruction effect: The scores for the memory training group and for the simple rule group were also significantly higher than the control group. See Figure 2.

The model setup. The model was set up as described earlier. We used 168 input units, 40 hidden units, and 12 output units. There were 7 groups of input units, each for a particular (past) time step, constituting a moving time window. Each group of input units contained 24 units, in which half of them encoded 12 system output levels and the other half encoded 12 system input levels at a particular step. The 12 output units indicated 12 levels of subjects' input to the system. The learning rate was 0.1. The momentum was 0.1.

The rule deletion threshold was set at 0.15 for simulating control subjects. To capture the verbalization condition, the rule deletion threshold was raised to 0.35 (to encourage more rule learning activities). To capture the explicit instruction conditions, in the "memory training" condition, each of the 12 examples was wired up at the top level as simple rules (in the form of $P_1 - \quad W$); in the "simple rule" condition, the simple rule (as described earlier) was wired up at the top level. A reward of 1 was given when the system output was within the target range. In simulating the person task (a common, everyday task), we used pre-training of 10 blocks before data collection, to capture prior knowledge subjects likely had in this type of task.

The match. Our simulation captured the verbalization effect in the human data well. See Figures 2 and 3. We used a t test to compare the "original" group with the control group in the model data, which showed a significant improvement of the original group over the control group ($p <$ 01), the same as the human data.

Our simulation also captured the explicit instruction effect, as shown in Figure 3. We used pair-wise t tests to compare the "memory training" and "simple rule" groups with the control group in the model data, which showed significant improvements of these two groups over the control group, respectively ($p <$ 01).

Both effects point to the positive role of the top level. When the top level is enhanced, either through verbalization or through externally given explicit instructions, performance is improved, although such improvement is not universal (Sun et al 2001). They both showed synergy between the top-level explicit processes and the bottom-level implicit processes.

Simulating Berry and Broadbent (1988)

The task. The task was similar to the computer "person" task in Stanley et al (1989). Subjects were to interact with a computer simulated "person" whose behavior ranged from "very rude" to "loving" and the task was to maintain the behavior at "very friendly" by controlling his/her own behavior (which could also range from "very rude" to "loving"). In the salient version of the task, the behavior of the computer "person" was determined by the immediately preceding input of the subject: It was usually two levels lower than the input ($P = W - 2 \quad N$). In the non-salient version, it was determined by the input before that and was again two levels lower than that input ($P = W_1 - 2 \quad N$). Noise (N) was added to the output of the computer "person" so that there was a chance of being up or down one level (a 33% chance respectively).

Four groups of subjects were used: salient experimental, salient control, non-salient experimental, and non-salient control. The experimental groups were given explicit search instructions after the first set of 20 trials, and after the second set of 20 trials were given explicit instructions in the form of indicating the relevant input that determined the computer responses (W or W_1). 12 subjects per group were tested.

The data. The exact target value plus/minus one level (that is, "friendly", "very friendly", or "affectionate") was considered on target. The average number of trials on target was recorded for each subject for each set of 20 trials. Figure 4 shows the data for the four groups of subjects for the three sets of trials. Analysis showed that on the first set, neither of the two experimental groups differed significantly from their respective control groups.

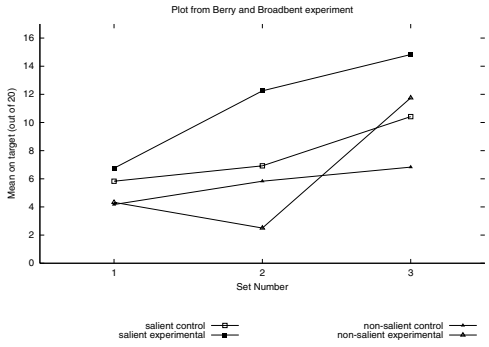


Figure 4: The data of Berry and Broadbent (1988).

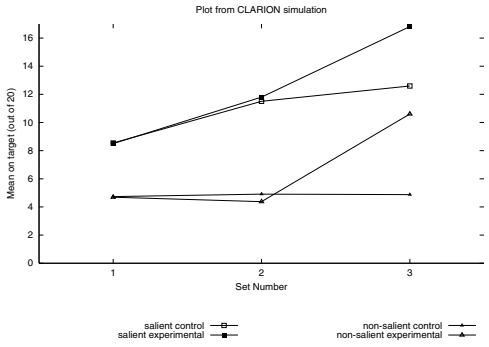


Figure 5: The simulation of Berry and Broadbent (1988).

However, on the second set, the salient experimental group scored significantly higher than the salient control group ($p < 0.01$), but the non-salient experimental group scored significantly less than the non-salient control group ($p < 0.05$). On the third set, both experimental groups scored significantly higher than their respective control groups ($p < 0.01$). The data clearly showed (1) the explicit search effect: improving performance in the salient condition and worsening performance in the non-salient condition; (2) the explicit instruction effect: improving performance in all conditions; as well as (3) the salience difference effect (during the 2nd set, under the explicit search condition).

The model setup. The model was set up similarly as described earlier for simulating Stanley et al (1989), except the following differences. The rule deletion threshold was set at 0.1 initially. To capture the explicit search effect (during the second training set), the rule deletion threshold was raised to 0.5 (for increased learning activities in the top level), and the weighting of the two levels was changed to 0.5/0.5 (for more reliance on the top level). To capture the explicit instructions given in this task (during the third training set), only rules that related the given critical variable to the system output were hypothesized and tested at the top level thereafter, in correspondence with the instructions (that is, $P = aW \quad b$, where W is the critical variable indicated by the instructions). The learning rate was 0.04. The momentum was 0.

The match. We captured in our simulation of this task the following effects exhibited in the human data: the salience difference effect, the explicit search effect, and the explicit instruction effect. The results of the simulation are shown in Figure 5. On the first set, neither of the two experimental groups differed significantly from their respective control groups; however, on the second set, the salient experimental group scored slightly higher than the salient control group, but the non-salient experimental group scored slightly less than the non-salient control group. On the third set, both experimental groups scored significantly higher than their respective control groups ($p < 0.01$).

The data demonstrated clearly the explicit instruction effect (improving performance in all conditions), and showed to some extent the explicit search effect (improving performance in the salient condition and worsening performance in the non-salient condition), as well as the salience difference effect along with the explicit search effect. The data showed the extent and the limit of the synergy effect (in that the non-salient condition discouraged synergy).

General Discussions

Although implicit learning is a controversial topic, the existence of implicit processes in skill learning is not in question — what is in question is their extent and importance. We allow for the possibility that both types of processes and both types of knowledge coexist and interact with each other to shape learning and performance, so we go beyond the controversies and the studies that focused mostly on the minute details of implicit learning (Gibson et al 1997).

The incorporation of both processes allows us to ask the question of how synergy is generated between the two separate, interacting components of the mind (the two types of processes). The model may shed some light on this issue. Sun and Peterson (1998) did a thorough computational analysis of the source of the synergy between the two levels of CLARION in learning and in performance. The conclusion, based on the systematic anaylsis, was that the explanation of the synergy between the two levels rests on the following factors: (1) the complementary representations of the two levels: discrete vs. continuous; (2) the complementary learning processes: one-shot rule learning vs. gradual Q-value approximation; and (3) the bottom-up rule learning criterion used in CLARION. [4] It is very likely, in view of the match between the model and human data as detailed in this paper, that the corresponding synergy in human performance results also from these same factors (in the main).

[4]Due to lengths, we will not repeat the analysis here. See Sun and Peterson (1998) for details.

As a result of its distinct emphasis, CLARION is clearly distinguishable from existing unified theories/architectures of cognition, such as SOAR, ACT, and EPIC. For example, SOAR (Rosenbloom et al 1993) is different from CLARION, because SOAR makes no distinction between explicit and implicit learning, and is based on specialization, using only symbolic forms of knowledge. Although ACT (Anderson 1993) makes the distinction, it is different from CLARION because traditionally it focuses mainly on top-down learning (from declarative to procedural knowledge).

Concluding Remarks

This work highlights the importance of the interaction of implicit and explicit processes in skill learning. It captures the interaction through a model that includes both types of processes. This modeling work reveals something new in the existing data (cf. Gibson et al 1997, Lebiere et al 1998). The contribution of this model lies in capturing human data in skill learning through the interaction of the two types of processes, and also in demonstrating the computational feasibility and psychological plausibility of bottom-up learning (Sun et al 2001).

References

M. Ahlum-Heath and F. DiVesta, (1986). The effect of conscious controlled verbalization of a cognitive strategy on transfer in problem solving. *Memory and Cognition*. 14, 281-285.

J. R. Anderson, (1993). *Rules of the Mind*. Lawrence Erlbaum Associates. Hillsdale, NJ.

D. Berry and D. Broadbent, (1988). Interactive tasks and the implicit-explicit distinction. *British Journal of Psychology*. 79, 251-272.

A. Clark and A. Karmiloff-Smith, (1993). The cognizer's innards: a psychological and philosophical perspective on the development of thought. *Mind and Language*. 8 (4), 487-519.

A. Cleeremans, (1994). Attention and awareness in sequence learning. *Proc. of Cogntiive Science Society Annual Conference*, 330-335.

Z. Dienes and R. Fahey, (1995). The role of specific instances in controlling a dynamic system. *Journal of Experimental Psychology: Learning, Memory and Cognition*. 21, 848-862.

F. Gibson, M. Fichman, and D. Plaut, (1997). Learning in dynamic decision tasks: computational model and empirical evidence. *Organizational Behavior and Human Decision Processes*, 71 (1), 1-35.

A. Karmiloff-Smith, (1986). From meta-processes to conscious access: evidence from children's metalinguistic and repair data. *Cognition*. 23. 95-147.

S. Keele, R. Ivry, E. Hazeltine, U. Mayr, and H. Heuer, (1998). The cognitive and neural architecture of sequence representation. Technical report No.98-03, University of Oregon.

C. Lebiere, D. Wallach, and N. Taatgen, (1998). Implicit and explicit learning in ACT-R. *Proc. of ECCM'98*, pp.183-189. Nottingham University Press.

P. Lewicki, M. Czyzewska, and H. Hoffman, (1987). Unconscious acquisition of complex procedural knowledge. *Journal of Experimental Psychology: Learning, Memory and Cognition*. 13 (4), 523-530.

R. Nosofsky, T. Palmeri, and S. McKinley, (1994). Rule-plus-exception model of classification learning. *Psychological Review*. 101 (1), 53-79.

M. Rabinowitz and N. Goldberg, (1995). Evaluating the structure-process hypothesis. In: F. Weinert and W. Schneider, (eds.) *Memory Performance and Competencies*. Lawrence Erlbaum, Hillsdale, NJ.

A. Reber, (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*. 118 (3), 219-235.

D. Rumelhart, J. McClelland and the PDP Research Group, (1986). *Parallel Distributed Processing*. MIT Press, Cambridge, MA.

P. Rosenbloom, J. Laird, and A. Newell, (1993). *The SOAR papers: Research on Integrated Intelligence*. MIT Press, Cambridge, MA.

C. Seger, (1994). Implicit learning. *Psychological Bulletin*. 115 (2), 163-196.

P. Smolensky, (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11 (1), 1-74.

W. Stanley, R. Mathews, R. Buss, and S. Kotler-Cope, (1989). Insight without awareness. *Quarterly Journal of Experimental Psychology*. 41A (3), 553-577.

R. Sun, (1995). Robust reasoning: integrating rule-based and similarity-based reasoning. *Artificial Intelligence*. 75, 2. 241-296.

R. Sun, (1997). Learning, action, and consciousness: a hybrid approach towards modeling consciousness. *Neural Networks*, special issue on consciousness. 10 (7), pp.1317-1331.

R. Sun and T. Peterson, (1998). Autonomous learning of sequential tasks: experiments and analyses. *IEEE Transactions on Neural Networks*, Vol.9, No.6, pp.1217-1234.

R. Sun, E. Merrill, and T. Peterson, (2001). From implicit skill to explicit knowledge: a bottom-up model of skill learning. *Cognitive Science*.

C. Watkins, (1989). *Learning with Delayed Rewards*. Ph.D Thesis, Cambridge University, Cambridge, UK.

D. Willingham, M. Nissen, and P. Bullemer, (1989). On the development of procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 15, 1047-1060.

Preserved Implicit Learning on both the Serial Reaction Time Task and Artificial Grammar in Patients with Parkinson's Disease

Jared G. Smith (smithjare@scs.vuw.ac.nz)
School of Psychology, Victoria University of Wellington,
PO Box 600, Wellington, New Zealand

Richard J. Siegert (Richard.Siegert@vuw.ac.nz)
School of Psychology, Victoria University of Wellington,
PO Box 600, Wellington, New Zealand

John McDowall (john.mcdowall@vuw.ac.nz)
School of Psychology, Victoria University of Wellington,
PO Box 600, Wellington, New Zealand

David Abernethy (david.abernethy@clear.net.nz)
Wellington School of Medicine, University of Otago,
PO Box 600, Wellington, New Zealand

Abstract

Thirteen non-demented patients with Parkinson's disease (PD) were compared with age matched controls on two standard tests of implicit learning. A verbal version of the Serial Reaction Time (SRT) task was used to assess sequence learning and an artificial grammar (AG) task assessed perceptual learning. It was predicted that PD patients would show implicit learning on the AG task but not the SRT task, as motor sequence learning is thought to be reliant upon the basal ganglia which is damaged in PD. Patients with PD demonstrated implicit learning on both tasks. In light of these unexpected results the research on SRT learning in PD is reconsidered, and some possible explanations for the sometimes conflicting results of PD patient samples on the SRT task are considered. Factors which merit further study in this regard are: The degree to which the SRT task relies on overt motor responses; the effects of frontal lobe dysfunction upon implicit sequence learning; and the degree to which the illness itself has advanced.

Current theoretical accounts of human memory draw an important distinction between implicit and explicit learning processes (e.g., Squire, 1994; Squire & Zola, 1996). Explicit (or declarative) learning and memory is characterized by the acquisition and retrieval of information accompanied by awareness of the learned information and its influence. Implicit learning refers to similar acquisition without awareness of the learned information or its influence. Such learning occurs in situations and tasks whereby the ability to consciously or deliberately recall the episode in which learning took place, or to describe the rules underlying the task, typically fall well behind the level of performance. It is thought that explicit learning is dependent upon medial temporal lobe and diencephalic brain structures, while habit learning and implicit skill learning is closely associated with neostriatal structures such as the basal ganglia (Squire, 1994; Squire & Zola, 1996).

One striking characteristic of implicit learning has been its demonstrable robustness even in the face of quite major brain damage. For example, using the serial reaction time (SRT) task researchers have shown implicit learning to be preserved in normal ageing (Howard & Howard, 1989), Korsakoff patients (Nissen & Bullemer, 1987), closed head injury patients (McDowall & Martin, 1996) and Alzheimer's disease (Knopman & Nissen, 1987). However, the fact that the very brain structures thought to be closely associated with certain forms of implicit learning are the most impaired by Parkinson's disease (PD), makes PD of special interest for implicit learning researchers. The characteristic neuropathology of PD includes marked degeneration and atrophy of the basal ganglia and substantia nigra, particularly the caudate nucleus or the neostriatum (Knight, 1992). In the present study we were interested to compare the performance of PD patients with controls on two tests of implicit learning: the serial reaction time (SRT) task and the artificial grammar (AG) task.

In the SRT task participants respond as quickly as possible to the presentation of an asterisk on a computer monitor. The asterisk can appear at any one of several different locations and participants must respond by pressing a key which corresponds to the spatial location of the asterisk. Unknown to the participants, the location of the stimulus follows a sequence which is repeated over a number of trials. Sequence learning is assumed to occur when, over the course of successive trials, the reaction times (RT) of participants decrease significantly and when there is a significant increase in the RT of participants upon the

administration of a block of trials where the position of the asterisk is random. Of particular interest, is that while participants display significant learning over trials, they are often unaware that such learning has occurred and mostly unable to correctly report the actual sequence followed (Nissen & Bullemer, 1987).

In the first study of PD patients using the SRT task Ferraro, Balota and Connor (1993) reported that non-demented individuals with PD showed less sequence specific learning than healthy controls. Similarly, Pascual-Leone et al. (1993) found patients with PD acquired some SRT procedural knowledge although its degree was less than in healthy volunteers. However, perhaps the clearest evidence for an implicit visuomotor learning deficit in patients with basal ganglia dysfunction comes from a study by Jackson et al. (1995). The authors found no significant SRT learning in PD patients and concluded that the results suggest a role for the basal ganglia in SRT learning or the expression of serially-ordered action. Westwater et al. (1998) employed a verbal version of the SRT task, designed to minimize the influence of the motor symptoms of PD, and reported similar results. In summary, there is a growing number of studies suggesting that implicit learning in PD, at least as measured by the SRT task, is reduced or impaired in people with PD.

AG learning involves presenting participants with a set of rule-governed stimuli (typically cards consisting of letter strings belonging to a finite-state grammar) for observation, and asking them to commit the letter strings to memory. The set of stimuli typically consists of exemplars which cover the entire range of transitions of the grammar, providing exposure to all the rules of the grammar albeit in an indirect fashion. On completion of the orientation task, participants are informed of the existence of a complex grammatical system governing the stimuli presented. Participants are then shown a new set of cards, only half of which conform to the grammar, and asked to decide whether each item conforms to the structure of the grammar. The assumption behind this paradigm is that tacit knowledge, which is abstract and representative of a complex grammar system, can be learnt independently of conscious efforts (Reber, 1989).

One important theoretical issue for the study of implicit learning concerns the degree to which different types of implicit learning are separate or dissociable both functionally and at an anatomical level. For instance, while formally similar to habit learning paradigms such as the SRT task, AG participants typically evidence abstract knowledge about a complex rule system on grammaticality tests, while the measure used in SRT tasks is reaction time, which is more likely to tap visuomotor knowledge (Seger, 1998). The examination of abstract judgment-linked learning (e.g., AG learning) and visuomotor learning in a group such as PD patients, where brain structures assumed to be involved in implicit learning processes are damaged, provides a method to investigate the possibility that these forms of implicit learning may be independent.

While several studies have examined the performance of PD patients on the SRT task (and generally found deficits or impairments), to our knowledge only two published studies have reported using the AG task with PD patients. Thierry, Peigneux and Van der Linen (1998) observed the same level of performance in controls and patients with PD on initial trials which suggested preserved AG learning in PD, and more broadly, that the basal ganglia may not be crucially involved in the rule-extraction mechanisms engaged in AG learning. Recently, Reber and Squire (1999) investigated the ability of patients with PD to learn AG in both a standard condition and a letterset transfer version of the task. They observed learning under both conditions and concluded that the learning of AGs appeared not to depend on the integrity of the neostriatum. They also commented that the dissociation between SRT and AG performance in patients with PD relies upon comparisons across studies, and that a dissociation within the same group of patients would be even stronger evidence.

The finding that patients with PD exhibit intact AG learning but show impairment on SRT tasks suggests implicit learning is not a single entity and that different neural systems may mediate performance on particular implicit learning tasks. In the present study we set out to compare the performance of a group of patients with PD on the SRT task with their performance on an AG task. The verbal version of the SRT task replaced the standard button-pressing response with a vocal response in an attempt to reduce the motor component of the task. We hypothesized that patients with PD would show impaired performance on the SRT task but not on the AG task in comparison to healthy controls.

Method

Participants

Participants consisted of 14 patients with PD recruited from the Neurology outpatients' service of Wellington Hospital, and 14 volunteers from the community who served as healthy controls. The diagnosis of PD was confirmed by a senior staff neurologist. One member of the PD group scored below the standard cut-off of 24 points on the Mini-Mental Status Examination (MMSE; Folstein, Folstein, & McHugh, 1975), used as a screening measure for abnormal cognitive decline, and was excluded from further analyses.

The PD group comprised eight males and five females, with a mean age of 66.42 years (range = 37 to 79 years). In the control group, eight were male and six were female, and the mean age was 68.36 years (range = 53 to 74 years). Each of the patients with PD in the present study fell within the early to middle/late stages of severity as assessed by the Hoehn and Yahr (1967) degree of clinical disability scale. Ten of the patients were in Stage Two (bilateral midline involvement without loss of balance), two were in Stage Three (first signs of impairment in equilibrium, significant slowing of body movements), and one was in Stage Four (fully developed PD, still able to stand and walk, but markedly incapacitated). At the time of testing all patients with PD were under the care of a neurologist and all but two were receiving anti-Parkinsonian medication. None had a history of head injury within the preceding ten years, or had a history of alcohol abuse, stroke or epilepsy, and all

subjects had normal or corrected to normal vision. The administration of a standardized measure of depression indicated an absence of depression for all participants.

Materials

All participants were administered the National Adult Reading Test (NART; Nelson & Willeson, 1991) to compare performance on intellectual ability. Additionally, the Controlled Oral Word Association Test (COWAT; Benton & Hamsher, 1976) was administered in order to assess verbal fluency. There were no significant group differences on variables of age, gender, or number of years spent in formal education. A summary of the group demographics is displayed in Table 1.

Table 1. Demographic Data.

Measure	PD M (SD)	Controls M (SD)
Age (yrs)	66.4 (11.0)	68.3 (8.4)
Education (yrs)	12.2 (2.7)	12.3 (2.6)
MMSE	27.3 (2.0)	29.0 (1.3)
COWAT	33.6 (12.2)	47.6 (2.1)*
NART	116.5 (5.0)	122.0(10)*

* $p<.05$

Note: MMSE = Mini Mental Status Examination; COWAT = Controlled Oral Word Association Test - age corrected scores: NART = National Adult Reading Test, expressed as a Wechler Adult Intelligence Scale - Revised full scale equivalent.

Apparatus and Procedure

All participants were tested individually beginning with the NART, followed by the MMSE. Following this participants either completed the SRT, the COWAT, and then the AG task, or performed these three tasks in reverse order. The ordering of these three tasks was counterbalanced within both the PD patient group and the control participant group.

SRT Task. The SRT task was a verbal version of the classic SRT task, as devised by Nissen and Bullemer (1987), replicating the SRT task used by Westwater et al. (1998) (refer to Westwater et al. for a more detailed description of the procedure). Briefly, all participants completed five blocks of trials, each consisting of 100 trials. In each trial a stimulus (an asterisk) appeared in one of four positions along the bottom of a computer monitor. In the first four blocks the asterisk appeared in a sequential manner (the 10-item sequence used in Nissen and Bullemer (1987)). In the fifth block the location of the asterisk was determined pseudorandomly. All participants were asked to respond as quickly as possible to the location of each stimulus by saying aloud the number corresponding to its location. Upon a response the stimulus disappeared and 400ms later the next stimulus appeared in one of the other locations. At the conclusion of the task all participants were asked whether they noticed anything about the nature of the stimuli. Although some participants reported being aware of some form of pattern to the stimuli, none were able to correctly reproduce it when asked to do so.

Artificial grammar task. Grammatical letter strings were generated from a finite-state Markovian rule system identical to that used by Dienes, Broadbent, and Berry (1991). This structure was used to generate both 23 training and 23 test items, each three to six letters in length. Twenty-three non-grammatical test items were also generated from the rule system by substituting an inappropriate letter for an appropriate letter in an otherwise grammatical string. Each letter string was presented on a 7.5 x 12.7 cm index card.

The procedure for the training and testing phases closely followed the standard AG procedure and is fully described by Dienes et al. (1991: Experiment 1., "grammatical" participants). At the conclusion of the task, participants were asked: "What were the grammatical rules or strategies on which you were basing your judgments of grammaticality or classification". No participant was able to accurately identify the rules with any significant success.

Results

SRT task. The results of one patient with PD were omitted from SRT data analyses because of a technical problem with the microphone and the voicebox, which led to invalid data. Error rates were defined as verbal responses which were incorrect with regard to the position of the stimuli, as well as any omissions. Both groups averaged well below a 5% error rate across blocks and did not differ significantly in total error rate, $t(24) = -1.22$, $p>.05$. Incorrect responses were not included in the RT analyses. For each set of 10 trials (the sequence pattern in blocks 1 to 4), each participant's median RT of correct responses was computed. Figure 1. shows the mean of those median scores for each block (ten repetitions of 10 trials) for the PD and control groups. All analyses involved a mixed Group x Block ANOVA with Block as a within-group factor. A 2 (Group) x 5 (Block) mixed factor ANOVA showed a significant Group effect, $F(1,24) = 6.34$, $p<.05$, and a significant effect of Block $F(4,96) = 11.75$, $p<.0001$. There was no significant Group x Block interaction, $F(4,96) = 0.56$, $p>.05$.

In order to examine both sequence learning and non-specific practice effects a 2 (Group) x 4 (Block) ANOVA with repeated measures on the last factor was computed over the first four blocks. This revealed a significant main effect for Block, $F(3,72) = 15.89$, $p<.0001$, and a significant main effect for Group, $F(1,24) = 6.02$, $p<.05$. There was no Group x Block interaction, $F(3,72) = 0.45$, $p>.05$.

Decreased RT over the first four blocks can result from both sequence learning and non-specific practice effects. To examine sequence-specific learning, a 2 (Group) by 2 (Block) mixed factor ANOVA was computed for Block 4 and Block 5. This resulted in main effects for Group, $F(1,24) = 7.72$, $p<.05$, and Block, $F(1,24) = 24.21$, $p<.0001$. There was no Group x Block interaction, $F(1,24) = 0.01$, $p>.05$.

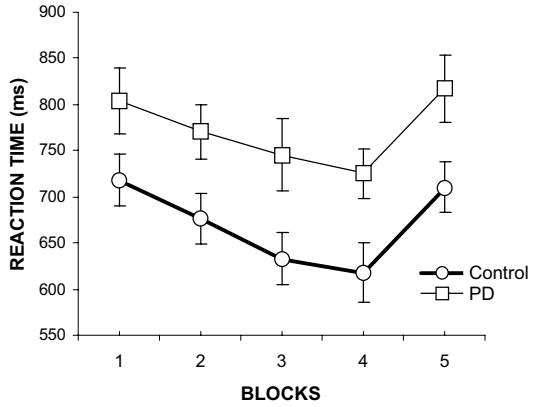


Figure 1. Mean RT across blocks for PD and Control groups.

A preliminary inspection of the effects of disease severity on sequence learning was performed by dividing patients into two groups according to their Hoehn and Yahr scores: Stage 2 (n=9) and Stage 3-4 (n=3). However, the data suggested there was no effect of severity of disease on SRT performance. These results have to be interpreted with caution because of the small number of patients especially in the severe group. A similar pattern of results emerged using both the COWAT and NART as covariates.

Frontal lobe dysfunction has also been associated with impairment in performance on a visuomotor sequence learning task (Beldarrain et al., 1999). However, correlation analysis failed to show a significant association between performance on the COWAT (a test that has been associated with frontal lobe functioning) and sequence specific learning (as measured by the increase in mean reaction time from trial 4 to 5) for patients with PD, $r = -0.56$, $p>.05$. Once again, small numbers preclude any serious conclusions on this matter.

Artificial grammar task. Participants' scores were calculated firstly, by the percentage of grammatical strings classified correctly and the percentage of ungrammatical strings classified correctly, and secondly, by the percentage of grammatical strings classified as grammatical relative to the percentage of grammatical strings classified as ungrammatical.

Average percentage correct for making grammaticality judgments for the patients with PD was 55.9% (standard error of the mean (SEM) = 2.1%), a performance significantly better than chance, $t(12) = 2.85$, $p<.01$. Controls obtained 57.9% (SEM = 1.9%) correct for grammaticality judgments, and also performed better than chance, $t(13) = 4.32$, $p<.001$. There was no significant difference in classification performance between the groups, $t(25) = 0.75$, $p>.05$.

Patients with PD classified as grammatical 61.2% (SEM = 3.4%) of the grammatical strings and 49.5% (SEM = 4.7%) of the ungrammatical strings. Control participants classified as grammatical 63.4% (SEM = 3.9%) of the grammatical strings and 47.5% (SEM = 3.7%) of the ungrammatical strings. A two-way ANOVA revealed a significant effect of the Grammaticality variable, $F(1,25)= 25.24$, $p<.0001$, but no significant Group effect, $F(1,25)= 0.0003$, $p>.05$. There was no significant Group x Grammaticality variable interaction, $F(1,25)= 0.57$, $p>.05$.

Discussion

The present study compared the performance of patients with PD with matched controls on two distinct tests of implicit learning, a verbal version of the SRT task and an AG task. Contrary to our first hypothesis the participants with PD demonstrated implicit learning on the SRT task. As predicted, they also showed implicit learning on the AG task. The control group also demonstrated implicit learning on both tasks. These results are further testimony to the robustness of implicit learning in the face of both age (given the mean ages of both groups) and neurological damage. At the same time the failure to observe impaired learning on the SRT task, by the PD participants, is inconsistent with other recent studies (e.g., Jackson et al., 1995; Westwater et al., 1998).

Perhaps the first point to consider is that findings regarding implicit learning and PD have been quite diverse and sometimes conflicting. For example, findings of deficits in performance of patients with PD on rotor-pursuit tasks (Harrington et al., 1990; Heindel et al., 1989) and mirror reading skill acquisition tasks (Allain et al., 1995; Yamadori et al., 1996) are tempered by findings of preserved learning on both the former (Bondi & Kasniak, 1991), and the latter (Bondi & Kasniak, 1991; Harrington et al., 1990). Moreover, attempts to relate findings at a behavioral or cognitive level, with likely neuroanatomical substrates have also produced a complex picture. For example, some authors attribute performance deficits to the disrupted basal ganglia in PD, or argue for a more specific emphasis on brain stem structures of the basal ganglia such as the substantia nigra, or other basal nuclei including the caudate nucleus or the putamen (e.g., Doyon et al., 1997). Others attribute the primary role to impaired neuroanatomical circuitry in PD (e.g., Bondi & Kaszniak, 1991; Heindel et al., 1989; Taylor, Saint-Cyr, & Lang, 1986), or more specifically the “complex loop” (e.g., Bondi & Kaszniak, 1991), whereas some authors have emphasized the importance of disturbed striatofrontal or caudate outflow in PD (e.g., Saint-Cyr, Taylor, & Lang, 1988). In summary, research on implicit learning in PD has produced conflicting results and also a wide range of possible explanations at the anatomical level.

However, studies employing the SRT task have been generally more consistent. Jackson et al. (1995) reported impairments on a variant of the SRT task in a group of 10 non-demented PD patients compared with healthy controls. Pascual-Leone et al. (1993) reported that patients with PD “achieved procedural knowledge” on the SRT task but at a slower rate than healthy controls. Ferraro et al. (1993) concluded that “there does appear to be some breakdown in implicit learning in non-demented PD individuals....” (p.175). Doyon et al. (1997) observed an impairment late in the sequence acquisition process on a version of the SRT for PD patients with a bilateral striatal-dysfunction. Finally,

Westwater et al. (1998) using a verbal version of the SRT task found implicit learning was impaired in PD. In summary, the evidence that procedural learning is impaired in PD, at least as measured by the SRT task, is generally more consistent that for other dimensions of implicit learning. In light of the SRT studies reviewed above, it is interesting to speculate as to why the PD patients in the present study demonstrated preserved implicit learning.

One possible explanation for this discrepancy then concerns the verbal version of the SRT task adopted for this experiment. Specifically, the current investigation was structured as to minimize the extent to which deficits displayed by the patient group could be artifacts of bradykinesia, akinesia, and/or motor arrests (symptoms commonly associated with PD), rather than failure to demonstrate implicit learning per se. The present findings suggest that difficulties in executing a motor response may be responsible for the impairment in implicit learning of patients with PD, as gauged by the standard SRT tasks which include an overt motor component in the method of response (e.g., Ferraro et al., 1993, Jackson et al., 1995; Pascal-Leone et al., 1993). However, this line of thought must be viewed with some reservations. Firstly, the SRT task used here replicated that of Westwater et al. (1998) who obtained results which conflict with this study. Secondly, findings of impaired PD patient performance on habit learning tasks that do not include a motor component (Knowlton, Mangels, & Squire, 1996) strongly suggest that the neostriatum is important not just for motor learning but also for acquiring non-motor dispositions that depend on new associations. Finally, the verbal response retains a motor element in which the deficiency is a salient feature of PD. For instance, bradykinesia has been associated with inappropriate and/or lengthy hesitations and a softening of the voice (becoming less audible), often accompanied by monotonous and hurried speech sounds (Knight, 1992). Therefore, while the exclusion of an overt motor component in the SRT task is useful in light of the motor difficulties experienced by patients with PD, it is by itself unlikely to account for the unexpected preserved learning exhibited by patients with PD in the current investigation.

A second reason that could account for the inconsistent SRT performance of PD patient samples observed in studies involves the possible role played by the frontal lobes in visuomotor sequence learning. Jackson et al. (1995) reported evidence for a procedural learning deficit in PD patients on the SRT task. However, when they compared PD patients who scored poorly on the Wisconsin Card Sorting Test (WCST) (suggesting a degree of frontal lobe dysfunction), with patients who scored normally on this test, the "frontal" group appeared to perform considerably worse than either the "non-frontal" group or the healthy controls. Unfortunately, their small sample size (11 PD patients) precluded a meaningful statistical comparison of these sub-groups. Beldarrain et al. (1999) examined SRT learning in 22 (non-PD) patients with unilateral prefrontal lesions and observed that learning was impaired in patients with lesions greater than 2cm in diameter. In concluding they argued for the "crucial role of the prefrontal cortex in procedural implicit learning" (p.1859). By contrast, Doyon et al. (1997), who studied PD patients specifically, concluded that implicit learning depended upon the "integrity of both the striatum and the cerebellum, but not of the frontal lobes" (p.219). In the present study, pairwise correlations between COWAT performance and sequence specific learning on the SRT failed to reach significance supporting Doyon and colleagues' findings. However, these results must be interpreted with caution given that numbers were small and the COWAT is a measure of verbal fluency and not frontal lobe integrity per se. Future research would be advised to adopt more precise measures of frontal lobe functioning such as the WCST. In summary, there is some evidence, although far from unequivocal, that the intact functioning of the prefrontal cortex may be important for procedural learning. If this can be substantiated, then it has obvious relevance for clarifying the performance of PD patients on the SRT task, given that "frontal dysfunction" is such a common symptom of PD (Taylor et al., 1986).

A third possible explanation for the inconsistent findings in this area concerns the stage of the disease. Presumably, if the implicit learning deficit is related to damage to the basal ganglia, then this will become increasingly obvious as the disease advances. In support of this Doyon et al. (1997) found that "only PD patients in more advanced stages of the disease showed an impairment in acquiring the repeating sequence" (p.235). Similarly, on the rotary pursuit task, also an example of implicit motor skill learning, Harrington et al. (1990) reported that procedural learning was impaired but only in patients with more advanced symptoms of PD. Interestingly, in the present study, a preliminary analysis of severity did not show any effect, although as 9 of the 12 patients were in Stage 2 on the Hoehn and Yahr scale, this is perhaps not surprising. Another, preferably continuous, measure of motor function or severity that allowed for a more even distribution of the subjects into two groups would have perhaps been more useful given the small number of patients.

Finally, it is important to note that implicit learning on the AG task was also preserved among the PD patients. Overall, both groups classified strings according to their grammatical status at a level above chance, demonstrating learning for the AG system, learning that could not be consciously articulated by participants in either group. These findings are consistent with those of both Reber and Squire (1999) and Thierry et al. (1998) who also observed preserved AG learning in patients with PD and is in accord with current notions that such learning of perceptual knowledge is more cortically mediated and less reliant upon subcortical structures (Reber & Squire, 1999). Though we have devoted most of the discussion to considering explanations for the unpredicted results on the SRT task, the results on the AG task are also important, as this is only the third published study to date reporting preserved implicit learning on this task in PD patients. As such it adds to the growing body of evidence for the robustness of this dimension of implicit learning even in the face of neurological illness.

References

Allain, H., Lieury, A., Quemener, V., Thomas, V., Reymann, J., & Gandon, J. (1995). Procedural memory and Parkinson's disease. *Dementia, 6,* 174-178.

Beldarrain, M. G., Grafman, J., Pascual-Leone, A., & Garcia-Monco, J. C. (1999). Procedural learning is impaired in patients with prefrontal lesions. *Neurology, 52,* 1853-1860.

Benton, A. L., & Hamsher, K. (1976). *Multilingual aphasia examination.* Iowa City: University of Iowa.

Bondi, M. W., & Kasniak, A. W. (1991). Implicit and explicit memory in Alzheimer's disease and Parkinson's disease. *Journal of Clinical and Experimental Neuropsychology, 13,* 339-358.

Dienes, Z., Broadbent, D., & Berry, D. (1991). Implicit and explicit knowledge bases in AG learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 875-887.

Doyon, J., Gaudreau, D., Laforce, R., Castonguay, M., Bedard, P. J., Bedard, F., & Bouchard, J-P. (1997). Role of the striatum, cerebellum, and frontal lobes in the learning of a visuomotor sequence. *Brain and Cognition, 34,* 218-245.

Ferraro, F. R., Balota, D. A., & Connor, L. T. (1993). Implicit memory and the formation of new associations in nondemented Parkinson's disease individuals and individuals with senile dementia of the Alzheimer's type: A serial reaction time investigation. *Brain and Cognition, 21,* 163-180.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-Mental State": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12,* 189-198.

Harrington, D. L., Haaland, K. Y., Yeo, R. A., & Marder, E. (1990). Procedural memory in Parkinson's disease: Impaired motor but not visuo-perceptual learning. *Journal of Clinical and Experimental Neuropsychology, 12,* 323-339.

Heindel, W. C., Salmon, D., Shults, C., Walicke, P. A., & Butters, N. (1989). Neuropsychological evidence for multiple implicit memory systems: A comparison of Alzheimer's, Huntington's and Parkinson's disease patients. *Journal of Neuroscience, 9,* 582-587.

Hoehn, M. M., & Yahr, M. D. (1967). Parkinsonism: Onset, progression, and mortality. *Neurology, 17,* 427-442.

Howard, D. V., & Howard, J. H., Jr. (1989). Age differences in learning serial patterns: Direct versus indirect measures. *Psychology and Aging, 4,* 357-364.

Jackson, G. M., Jackson, S. R., Harrison, J., Henderson, L., & Kennard, C. (1995). Serial reaction time learning and Parkinson's disease: Evidence for a procedural learning deficit. *Neuropsychologia, 33,* 577-593.

Knight, R. G. (1992). *The neuropsychology of degenerative disorders.* Hillsdale, NJ: Lawrence Erlbaum.

Knopman, D., & Nissen, M. J. (1987). Implicit learning in patients with probable Alzheimer's disease. *Neurology, 37,* 784-788.

Knowlton, B. J., Mangels J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science, 273,* 1399-1402.

McDowall, J. & Martin, S. (1996). Implicit learning in closed-head-injured subjects: Evidence from an event sequence learning task. *New Zealand Journal of Psychology, 25,* 2-6.

Nelson, H. E., & Willeson, J. R. (1991). *National Adult Reading Test* (2nd ed.). Windsor, UK: NFER-Nelson.

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology, 19,* 1-32.

Pascual-Leone, A., Grafman, J., Clark, K., Stewart, M., Massaquoi, S., Lou, J., & Hallett, M. (1993). Procedural learning in Parkinson's disease and cerebellar degeneration. *Annals of Neurology, 34,* 594-602.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General, 118,* 219-235.

Reber, P. F., & Squire, L. R. (1999). Intact learning of AGs and intact category learning by patients with Parkinson's disease. *Behavioral Neuroscience, 113,* 235-242.

Saint-Cyr, J. A., Taylor, A. E., & Lang, A. E. (1988). Procedural learning and neostriatal dysfunction in man. *Brain, 111,* 941-959.

Seger, C. A. (1998). Multiple forms of implicit learning. In M. A. Stadler, and P. A. Frensch (Eds.), *Handbook of implicit learning.* London: Sage.

Squire, L. R. (1994). Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. In D. L. Schacter, and E. Tulving (Eds.), *Memory systems*. Cambridge, Mass.: MIT Press.

Squire, L. R., & Zola, S. M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences, 93,* 13515-13522 .

Taylor, A. E., Saint-Cyr, J. A., & Lang, A. E. (1986). Frontal lobe dysfunction in Parkinson's disease. *Brain, 109,* 845-883.

Thierry, M., Peigneux, P., & Van der Linen, M. (1998). Preserved AG learning in Parkinson's disease. *Brain and Cognition, 37,* 109-112.

Westwater, H., McDowall, J., Siegert, R., Mossman, S., & Abernethy, D. (1998). Implicit learning in Parkinson's disease: Evidence from a verbal version of the serial reaction time task. *Journal of Clinical and Experimental Neuropsychology, 20,* 413-418.

Yamadori, A., Yoshida, T., Mori, E., & Yamashita, H. (1996). Neurological basis of skill learning. *Cognitive Brain Research, 5,* 49-54.

On choosing the parse with the scene:

The role of visual context and verb bias in ambiguity resolution

Jesse Snedeker (jessned@psych.upenn.edu)
Kirsten Thorpe (thorpe@psych.upenn.edu)
John Trueswell (trueswel@ psych.upenn.edu)
The Department of Psychology and the Institute for Research in Cognitive Science
Philadelphia, PA 19104 USA

Abstract

Two striking contrasts currently exist in the sentence processing literature. First, whereas adult reading studies indicate an important role for verb biases in the initial consideration of syntactic alternatives (Britt, 1994), studies of adult listeners in world-situated eye-gaze studies suggest an almost exclusive role for referential cues in determining initial syntactic choices (Tanenhaus et al., 1995). Second, in contrast to adults, children in similar listening studies fail to take into account this referential information and instead appear to rely exclusively on verb biases or perhaps syntactically-based parsing principles (Trueswell et al., 1999). The current paper seeks to understand better these surprising contrasts by fully crossing verb bias and referential manipulations within a single experimental design, while using the eye-gaze listening technique. The full pattern is examined in adults (Exp. 1) and children (Exp. 2). Results indicate that adults combine both verb bias and referential information to determine syntactic choice, but children rely exclusively on verb bias. We discuss the implications for current theories of sentence processing as well as prior interpretations of world-situated listening studies.

Introduction

A central interest in the study of human language comprehension has been to understand the role that context plays in resolving linguistic ambiguities. In particular, can readers and listeners take into account extra-sentential information (i.e., information about the current situation or discourse) when making initial decisions about how to structure an incoming utterance? Or, do constraints on the organization of the comprehension system force it to exclude these non-linguistic factors during the early stages of processing?

These questions have played themselves out in the sentence processing literature in a series of studies examining how the referential context of a sentence affects the way readers initially interpret syntactically ambiguous phrases. To illustrate these findings, consider sentence fragment 1. The prepositional phrase (PP) beginning with *with* is temporarily ambiguous because it could be linked to the verb *hit* (verb phrase (VP)-attachment), indicating an Instrument (e.g., *with the stick*); or it could be linked to the definite noun phrase *the thief* (noun phrase (NP)-attachment) indicating a Modifier (e.g., *with the wart*).

1. The store owner hit the thief with the…

Crain and Steedman (1985) hypothesized that ambiguities involving this structure, and others, are initially resolved by taking into account the referential presuppositions of the syntactic analyses, with readers pursuing the analysis that has the fewest presuppositions. In short, if one assumes that a definite NP like *the thief* requires a unique referent, a restrictive modifier analysis of *with the wart* would presuppose the presence of two or more thieves, one of which has a wart. An Instrument analysis makes no such presupposition. Hence, it is predicted that in a context containing two possible referents (two-referent contexts) readers should pursue a modifier (NP-attachment) analysis, but in a one-referent context, or even a null context, readers should erroneously pursue an Instrument (VP-attachment) analysis.

Indeed several studies have found that readers in a two-referent context pursue a modifier analysis for ambiguous phrases of this sort (e.g., Altmann & Steedman, 1988; van Berkum, Brown & Hagoort, 1999, among many). However, several studies have failed to find such affects (e.g., Ferreira & Clifton, 1986; Rayner, Garrod & Perfetti, 1992).

An account of these conflicting findings comes from constraint-satisfaction theories of parsing that propose a role for verb biases in parsing preferences (e.g., MacDonald et al., 1994; Trueswell & Tanenhaus, 1994). These theories predict that referential effects should be weakened or eliminated when lexically specific constraints are strong. Thus, differences in the materials that were used in these prior studies may account for the conflicting findings.

Indeed, studies that have manipulated both referential context and verb bias have found that effects of referential factors disappear when a verb strongly prefers a single analysis (e.g., Britt, 1994; Spivey-Knowlton & Sedivy, 1995). Using materials like "*Susan put/dropped the book on the civil war onto the table"* Britt (1994) found that 2-book vs. 1-book contexts failed to guide parsing preferences when the verb required a PP argument. That is, for verbs like *put*, readers initially pursued VP-attachment regardless of context but for verbs like *dropped*, context guided parsing. These reading studies suggest that context only has an influence in the absence of strong lexical constraints, leading some researchers to contend that verb information plays the privileged role of proposing syntactic structures, which are only compared against context at a later stage (Boland & Cutler, 1996; Britt, 1994).

Recent work on syntactic ambiguity resolution in spoken language comprehension however has raised questions about the relative contributions of context and verb information. Tanenhaus, Spivey and colleagues (Tanenhaus et al., 1995;

Spivey et al., 2001) have found that under the right conditions situation-specific contextual information can completely override strong verb biases that support a competing syntactic alternative. In their studies, participants were given spoken instructions to move objects about on a table while their eye movements were recorded. Target instructions, like 2 below, contained a temporary PP-attachment ambiguity, in which the verb's argument preferences strongly supported an initial VP-attachment analysis of *on the napkin*.

2. Put the apple on the napkin into the box.

Even though the verb *put* requires a destination role, usually a PP, the two-referent context was sufficient to allow listeners to override the strong bias for VP-attachment. In particular, scenes containing two apples, one of which was on a napkin, eliminated early and late looks to an incorrect destination object (e.g., an empty napkin). Similar scenes with one apple resulted in large numbers of early and late looks to the incorrect destination. The authors concluded that when referential cues to attachment are salient, co-present with the linguistic utterance, and hence easy to maintain in memory, they can prevail over even the strongest of verb biases. However, they also noted that such strong effects of context are unexpected under most views of constraint-satisfaction, given the overwhelming structural bias of *put*.

Trueswell, Sekerina, Hill & Logrip (1999) replicated the findings of Tanenhaus et al. (1995) using essentially the same auditory eye-gaze task. In addition, they tested children, ages 4 and 5 , with the same materials. The children pursued the VP-attachment analysis, ignoring referential constraints even for the purpose of reanalysis. In particular, both two-referent and one-referent scenes showed early and late eye movements to the incorrect destination. Moreover, children's actions frequently involved the incorrect destination (e.g., moving an object to the empty napkin). By age eight, children acted like adults in this task, using referential context to guide parsing commitments. The authors concluded that the child parsing system relies heavily on verb-argument preferences to assign structure, and that processing demands prevented any use of the referential facts.[1] This developmental shift is surprising and a bit mysterious. How and why would lexicalist children become referentially-driven adults?

The current paper explores the striking and somewhat puzzling contrasts that we have outlined above. First, we wish to better understand the differences between adult reading and auditory studies, which paradoxically suggest that verb-specific preferences play little or no role in world-situated syntactic ambiguity resolution. Second, we wish to better understand the developmental change that occurs in sentence processing, to discover whether the parsing strategies of children and adults are as incommensurable as they appear.

To achieve these goals, we follow the lead of the prior reading studies that have, in a single experiment, fully crossed verb bias preferences with manipulations of referential context, except we now perform these manipulations in the world-situated eye-gaze task of Tanenhaus and colleagues. Such manipulations should reveal the relative contributions of these factors under all possible combinations. Second, we collected similar observations in five year olds, to observe the full pattern of information combination in this age group.

Experiment 1

In this experiment adults heard instruction containing a PP-attachment ambiguity (e.g., "Feel the frog with the feather") in both two-referent and one-referent contexts. For some subjects the target sentence contained a verb that typically uses an instrument phrase. Others heard instructions containing verbs that rarely use instruments. A third group was given instructions with equi-biased verbs.

The target instructions were globally ambiguous sentences rather than the temporarily ambiguous sentences typically used in comprehension studies. This was done for two reasons. First, we wanted to use the simplest sentences possible (i.e., ones without a second preposition) to avoid confusing children with uncommon sentence types (Exp. 2). Second, we were concerned that the previous listening studies (Tanenhaus et al., 1995; Trueswell et al., 1999) may have failed to find evidence that a VP-analysis was being considered in two-referent contexts because the disambiguating preposition occurred so soon after the introduction of the ambiguous phrase (see MacDonald, 1994, for the effects for post-ambiguity cues on parallel processing).

Because the sentences used in this study are never definitively disambiguated, we should expect continuity between the listeners' online attachment preferences and their ultimate interpretations. If listeners rely entirely on the visual context, then in two referent contexts they should interpret the ambiguous phrase as a modifier, regardless of verb bias. This preference should be reflected in both their eye movements and their actions. In contrast, if listeners simultaneously consider both lexical and contextual information then we would expect to find: 1) an effect of verb bias in both the one- and two-referent contexts and 2) an effect of referential context in some or all of the verb classes.

Methods

Participants Thirty-six students at the University of Pennsylvania volunteered for the experiment (twelve in each of the verb bias conditions). They received extra course credit or were paid for their participation. Twelve of the participants were males and all were native speakers of English.

Procedure The adult subjects were told that they were going to listen to and follow prerecorded instructions and that their responses would serve as a point of comparison for a study of how children follow directions. The subject sat in front of an inclined podium. At the center of the podium was a hole for a camera that focused on the subject's face. In each quadrant of the podium was a shelf where one of the props

[1] The children's parsing pattern might instead be attributable to the use of a syntactically-based parsing strategy (e.g., Minimal Attachment, Frazier & Fodor, 1978). This will be addressed in Experiment 2.

could be placed. At the beginning of each trial one experimenter laid out the props and introduced each one using indefinite noun phrases (e.g., *This bag contains a dog, a fan…*).

A second experimenter then played three prerecorded sound files from a laptop computer connected to external speakers. The first sound file was the same on every trial and simply told the subject to look at a fixation point at the center of the display. The second and third sound files were single sentence commands involving the props. The subject heard the first command, performed that action, and then heard the second command. Subjects signaled that an action was completed by saying "done". A second camera, placed behind the subject, recorded their actions and the locations of the props.

Stimuli On the critical trials, the first command contained an ambiguous Prepositional Phrase attachment, as in (3 a-c) below. The scene that accompanied these sentences contained the following objects: 1) a Target Instrument, a full scale object that could be used to carry out the action (e.g., for 3b a large feather); 2) a Target Animal, a stuffed animal carrying a small replica of the Target Instrument (e.g., a frog holding a little feather); 3) a Distractor Instrument; a second full scale object (e.g., a candle); and 4) A Distractor Animal, a stuffed animal carrying a replica of the Distractor Instrument. For Two Referent Trials the Distractor Animal and Target Animal were of the same kind (e.g., both frogs) for the One Referent Trials the Distractor Animal was of a different kind (e.g., a leopard carrying a candle).

3a. Choose the cow with the stick. (Modifier Bias)
3b. Feel the frog with the feather (Equi Bias)
3c. Tickle the pig with the fan. (Instrument Bias)

Examples of the three different types of verbs were used in this study are given in (3a-c). The verbs were identified in an earlier sentence completion study (see Snedeker, Dardick & Trueswell, 1999). In that experiment, adult subjects were asked to complete sentence fragments that ended with the ambiguously attached preposition (e.g., "Touch the teddy bear with…"). The verbs in the Modifier Bias condition were ones for which modifier completions (e.g., "the big brown eyes") were at least three times as frequent as Instrument completions (e.g., "your toes"). For the Instrument Bias verbs the opposite rule applied. Equi Bias verbs were those that fell somewhere in between.

The Target Instruments for each sentence were also chosen on the basis of a prior norming study (Snedeker et al., 1999). Subjects were shown several objects for each verb and asked to rate them as instruments for performing that action on a seven-point scale. We selected objects with mean ratings between 2 and 5 and balanced the ratings across the three Verb Bias conditions (M = 3.60, 3.65, and 3.64 for Modifier, Equi, and Instrument Biased respectively, $p > .9$).

Two presentation lists were constructed for each Verb Bias condition, so that each of the 8 target trials appeared in only one of the conditions on a given list but appeared in both conditions across lists (resulting in four target trials in each condition per subject). Thus Verb Bias was manipulated between subjects. This was done to minimize the number of trials per participant to ensure that children could complete the same study. Referential Context was manipulated within subjects but blocked. The first half of one list contained all One Referent Contexts while the first half of the other list contained just Two Referent Contexts. The critical trials were interspersed with twenty-four filler trials. The prop sets for the filler trials were similar to those used in the target trials: the attributes of the animals were matched to the large objects and animals of the same kind were used in half of the filler prop sets. Each list was presented in two orders (forward and reverse).

Coding Trained coders watched the videotape of the subject's actions and judged whether they made an Instrument response (performed the target action using the Target Instrument or the miniature instrument). A different coder viewed the videotape of the subject's face and recorded the onset of the target sentence and the onset and location of each fixation that occurred from the beginning of the instruction up until the subject began the action.

Results

Eye Movements For each trial we determined whether the subject looked at the Target Instrument during the time between the onset of the direct-object noun and the beginning of the action.[2] Figure 1 shows the proportion of trials with Instrument Fixations in each of the six conditions.

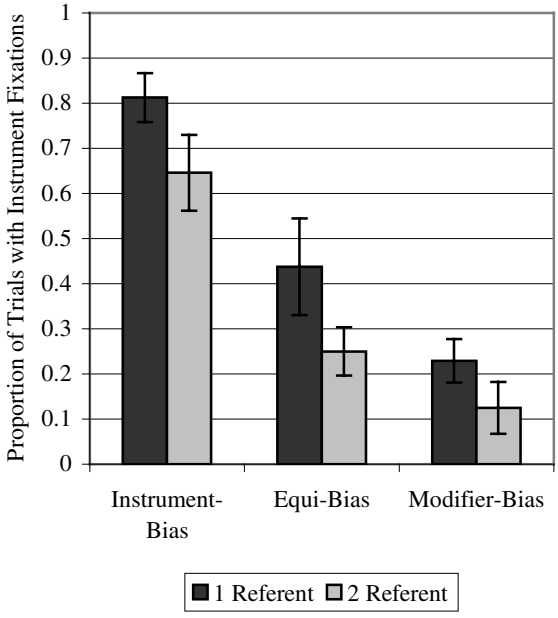


Figure 1: Proportion of Instrument Fixations for Adults (Experiment 1)

[2] This is essentially the same measure used by Tanenhaus et al., 1995. More fine-grained analyses of the pattern of fixations over time, indicate an early use of verb information even in Two Referent contexts.

Subjects' fixations during the ambiguous instructions were strongly affected by the type of verb in the sentence ($F1(2,24) = 27.71$, $p < .001$; $F2(2,18) = 35.01$, $p < .001$). Subjects who heard Instrument Biased verbs looked at the Target Instrument on 73% of the trials, indicating that they were considering the VP-attachment. Those who were given Modifier Biased verbs looked at the Target Instrument on only 18% of the trials.

Referential Context also had a strong and reliable effect on performance ($F1(1,24) = 10.52$, $p < .005$; $F2(1,18) = 11.90$, $p < .005$). When the ambiguous sentence occurred in a Two Referent Context only 34% of the trials included an Instrument Fixation, while in One Referent Contexts 49% of the trials did so. There was no significant interaction between Verb Type and Referential Context ($F1(2,24) > 1$, $p > .5$; $F2(2,18) > 1$, $p > .5$).

Actions The analysis of the Actions closely paralleled the analysis of the Instrument Fixations. Subjects tended to look at the Target Instrument when they were going to use it to perform the action but seldom fixated on it otherwise. The proportion of Instrument responses in each of the six conditions is presented in Figure 2.

Again there was a large and reliable effect of Verb Type ($F1(2,24) = 36.54$, $p < .001$; $F2(2,18) = 69.99$, $p < .001$). When the subjects heard an Instrument Biased verb, they produced Instrument actions 77% of the time. When they heard a Modifier Biased verb, they produced Instrument actions only 7% of the time.

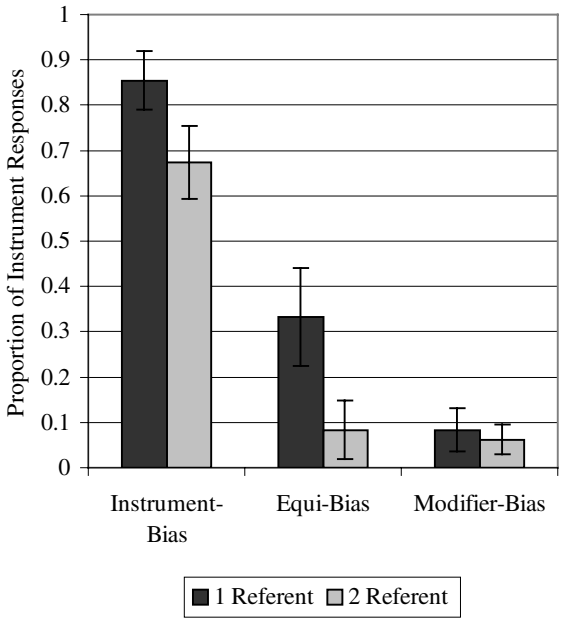


Figure 2: Proportion of Instrument Actions for Adults (Experiment 1).

Referential Context also had a strong effect on performance ($F1(1,24) = 10.81$, $p < .005$; $F2(1,18) = 15.99$, $p < .001$). In One Referent Contexts 42% of the responses involved the Target Instrument, in Two Referent Contexts only 27% did. Although the interaction between Verb Type and Referential Context was not reliable ($F1(2,24) = 2.20$, $p < .2$; $F2(2,18) = 3.28$, $p = .06$), the effect of context appeared to be isolated to the Equi Biased Verbs ($F1(1,8) = 5.33$, $p < .05$; $F2(1,6) = 11.39$, $p < .05$) and Instrument Biased Verbs ($F1(1,8) = 5.59$, $p < .05$; $F2(1,6) = 4.74$, $p = .07$). There was no significant effect of Referential Context for the Modifier Biased Verbs ($F1(1,8) < 1$, $p > .5$; $F2(1,6) = 1.00$, $p > .3$).

Experiment 2

A very similar experiment was conducted with five-year old children. Recall that Trueswell et al (1999) found an overwhelming VP-attachment bias in children of this age. As mentioned above this finding could be the result of the strong attachment bias of *put* or it could be evidence that children use a general structural parsing principle (e.g., minimal attachment). This experiment gives us the opportunity to distinguish between these explanations. A lexically based theory would predict that attachment preferences would be guided by verb information. A minimal attachment explanation would predict that children would show a VP-attachment preference independent of verb type. In addition, manipulating verb type allows us to see whether children's failure to use referential context is limited to strongly biased verbs (ala, Britt, 1994). We reasoned that children might prove to be sensitive to context for the Equi Biased verbs.

Methods

Participants Thirty-six children between 4;6 and 5;10 participated in the study ($M = 5;1$). Parents were contacted from Philadelphia area preschools and a commercial mailing list. Four additional children participated but were not included in the analyses because they refused to cooperate (1) were bilingual (1), or had been identified as developmentally delayed (2). Half of the children were male. Sex and age were balanced across the Verb Bias conditions and Lists.

Procedure and Stimuli The procedure was identical to Experiment 1 with the following exceptions. First, the children were told the names of each object twice. Second, the children were not asked to tell us when they had finished performing each action. Instead the experimenter who introduced the toys waited until the child finished moving the toys or looked at her and then praised the child for her response regardless of his or her action. Third, the number of filler trials was reduced from 24 to 10.

Results

Eye Movements Figure 3 shows the proportion of trials with Instrument Fixations in each of the six conditions. An Instrument Fixation was defined as any fixation to the Target Instrument that occurred between the onset of the direct object noun and the initiation of the action.

Like the adults, the children's fixations were strongly affected by the type of verb in the sentence ($F1(2,24) = 43.49$, $p < .001$; $F2(2,18) = 18.60$, $p < .001$). Subjects who heard Instrument Biased verbs looked to the Target Instrument on

82% of the trials, while those who heard the Modifier Biased verbs looked at the Target Instrument on only 21% of the trials. In contrast, Referential Context had no significant effect on the children's Instrument Fixations ($F1(1,24) < 1$, $p > .5$; $F2(1,18) = 1.41$, $p > .25$). There was no significant interaction between Verb Type and Referential Context ($F1(2,24) > 1$, $p > .5$; $F2(2,18) > 1$, $p > .5$). The children's fixations suggest that in all three Verb Bias Conditions, Referential Context played no role in determining the attachment of the ambiguous phrase.[3]

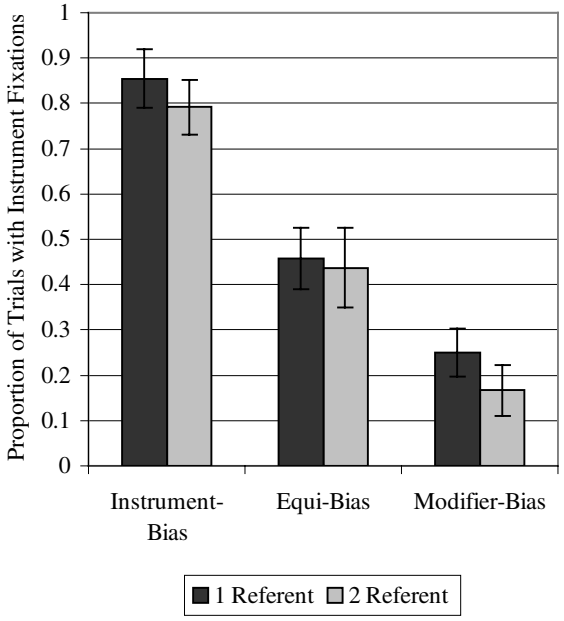
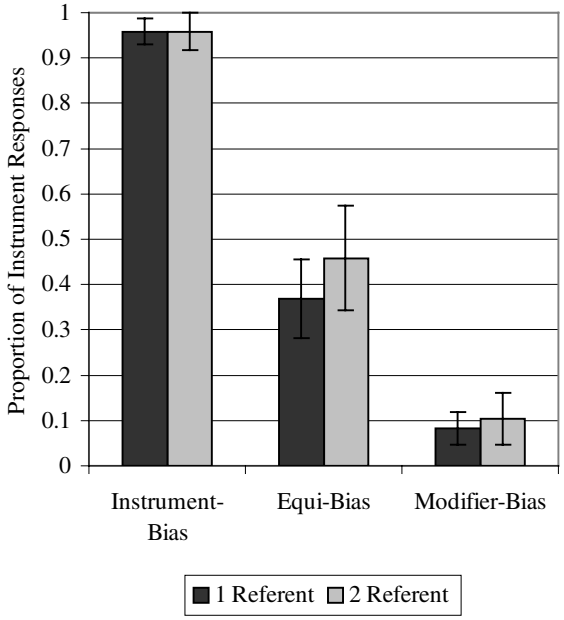


Figure 3: Proportion of Instrument Fixations for Five-Year Olds (Experiment 2)

Actions The proportion of Instrument responses in each condition is presented in Figure 4. Here again the offline actions and the online eye movements provide convergent evidence of the children's interpretation of the *with*-phrase.

Verb Bias had a striking effect on the children's actions ($F1(2,24) = 58.21$, $p < .001$; $F2(2,18) = 309.47$, $p < .001$). When the Subjects heard an Instrument Biased verb, they produced Instrument Actions 96% of the time. When they heard a Modifier Biased verb, they produced Instrument actions only 9% of the time. In contrast Referential Context appeared to have no effect on the children's actions ($F1(1,24) = 1.15$, $p > .2$; $F2(1,18) = 1.46$, $p > .2$). In One Referent Contexts instruments were used on 47% of the trials, in Two Referent Contexts they were used on 51% of the trials. There was no significant interaction between Referential Context and Verb Bias ($F1(2,24) < 1$, $p > .5$; $F2(2,18) < 1$, $p > .4$).

A direct comparison of the data from the two experiments revealed a main effect of Age Group ($F1(1,66) = 9.57$, $p < .005$; $F2(1,21) = 17.32$, $p < .001$) and an Age Group by Referential Context interaction ($F1(1,66) = 7.66$, $p < .01$; $F2(1,21) = 9.79$, $p < .005$). Five-year-old children produce



more Instrument responses than adults but this difference is limited to the Two Referent Condition.

Figure 4: Proportion of Instrument Actions for Five-Year Olds (Experiment 2)

In the One Referent Context, there was no effect of Age Group nor an Age Group by Verb Bias interaction (all Fs < 1, all p's > .3), indicating that the children and adults were equally responsive to the combinatorial properties of the verbs. In the Two Referent Contexts, there was both an effect of Age Group ($F1(1,66) = 16.54$, $p < .001$; $F2(1,21) = 26.93$, $p < .001$) and a marginal Age Group by Verb Bias interaction ($F1(2,66) = 3.00$, $p = .06$; $F2(2,21) = 4.90$, $p < .05$). Children gave more Instrument Responses, especially in the Instrument Biased and Equi Biased Conditions.

General Discussion

Two important findings emerge from this work. First, we observe that lexical biases do play an important role in adult parsing preferences in a world-situated task. Even when there is a rich and potentially constraining context that is co-present with the utterance, verb bias and referential cues combine to determine adult listeners' parsing preferences. Second, children show a complete inability to use referential information to inform parsing decisions, and instead reveal detailed sensitivity to verb biases. The implications of the adult and child data are considered separately below.

The data from our adult study indicate a greater continuity between the reading and listening than previous studies would suggest. Like Britt (1994), we observe contributions of both factors in on-line parsing commitments. This pattern is consistent either with a constraint-satisfaction approach that weighs both sources of evidence (e.g., Trueswell & Tanenhaus, 1994) or a 'propose-and-select' model which gives

[3]More detailed analyses of the time course of the eye movements support these claims. Verb Bias has an early effect on fixations but Referential Context does not.

a privileged status to lexical items in computing syntactic alternatives (e.g., Boland & Cutler, 1996).

Why then did the previous *put* studies show no consideration of the VP-attachment analysis? We speculate that two additional sources of information present in those studies may have further reduced consideration of VP-attachment. First, the appearance of a second prepositional phrase (*into the box*) right at the very moment that eye movements should show consideration of VP-attachment may have served as a post-ambiguity cue that squelched consideration of this parse. Second, prosodic cues may have provided evidence during the first PP that a second potential argument was forthcoming. Prosody was held constant across conditions in these studies, but the neutral prosody that the experimenters aimed for may have revealed that the utterance would continue. Indeed, our own studies of prosody, which used a similar task and measure, suggest that differences of this kind can influence parsing as rapidly as lexical information (Snedeker et al, 2001). These additional cues may not have been enough to completely eliminate the VP-attachment analysis in the one referent condition but may have been adequate to eliminate it in the two referent condition where context also supports a modifier analysis.

Implications from the child data are clear. First, children are not 'miniature minimal attachers'. The lack of a general VP-attachment bias, and a clear sensitivity to verb information speaks to this issue. Second, children seem instead to have formed parsing strategies that derive from their syntactic/semantic knowledge of individual verbs, lending further support to constraint-based lexicalist models of parsing.

An issue that remains less understood is why children fail to use referential specificity to guide their parsing commitments (i.e., the Referential Principle). This failure occurs even verbs that have no strong attachment preferences, which might override the effects of context. We strongly suspect that the failure to employ the referential principle is not due to a general lack of knowledge about specificity or the proper use of modification—our own studies show a clear talent in children's utterances for specifying a referent via locative modification (e.g., Hurewitz et al., 2001).

A controversial position, which our current data cannot rule out, is that children show a degree of bottom-up priority for lexically-based cues to syntax, perhaps because of the architectural configuration of the system. If children have memory limitations that prevent them from considering improbable syntactic alternatives, and probability is determined solely by distributional facts gleaned from utterances, then such a pattern might emerge. Only after the processing system gains the ability to maintain parallel parses over numerous words may the contextual facts further drive processing decisions. Indeed, this may also explain the inability of children in the Trueswell et al. study to revise initial commitments. It remains to be seen however, whether other contextual factors (e.g., related to conversational goals of a discourse) might better guide parsing preferences in children.

Acknowledgments

We thank Amy Nichols, Jessica Lilleston, John Paul Moorehead, Stefanie Poulos, Kate Ruane, Sandy Yim, and Lauren Cornew for their assistance with testing, coding and subject recruitment. We also gratefully acknowledge Tracy Dardick who carried out the norming studies. This work was supported by NIH Grant 1-R01-HD3750707-01 and a NSF Center Grant to the University of Pennsylvania Institute for Research in Cognitive Science.

References

Altmann, G. & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, 30, 191-238.

Boland, J. & Cutler, A. (1996). Interaction with autonomy: Multiple output models and the inadequacy of the Great Divide. *Cognition*, 58, 309-320.

Britt, M. A. (1994). The interaction of referential ambiguity and argument structure in the parsing of prepositional phrases. *JML*, 33, 251-283.

Crain, S. & Steedman, M. (1985). On not being led up the garden path: The use of context by the psychological parser. In Dowty, Karrattunen & Zwicky. *Natural Language Parsing,* Cambridge: Cambridge University Press.

Ferreira, F. & Clifton, C. (1986). The independence of syntactic processing. *JML*, 25, 348-368.

Frazier, L. & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, 6, 291-325.

Hurewitz, F., Brown-Schmidt, S., Thorpe, K., Gleitman, L., & Trueswell, J. (2000). One frog, two frog, red frog, blue frog: Factors affecting children's syntactic choices in production and comprehension. *JPR*, 29, 597-626.

MacDonald, M.C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *LCP*, 9, 157-201.

MacDonald, M.C., Pearlmutter, N.J., Seidenberg, M.S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101, 676-703.

Rayner, K, Garrod, S & Perfetti, C. (1992). Discourse influences during parsing are delayed. *Cognition*. 45, 109-139.

Snedeker, J., Dardick, T., & Trueswell, J. (1999). Identifying Verb Biases. Unpublished manuscript.

Spivey-Knowlton, M.J. & Sedivy, J. (1995) Resolving attachment ambiguities with multiple constraints. *Cognition*, **55**, 227-267.

Spivey, M.J., Tanenhaus, M.K., Eberhard, K.M. & Sedivy, J.C. (2001). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. In press, *Cognitive Psychology*.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632-1634.

Trueswell, J., Sekerina, I., Hill, N., & Logrip, M. (1999). The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition*, 73, 89-134.

Trueswell, J.C. & Tanenhaus, M. K. (1994). Toward a lexicalist framework of constraint-based syntactic ambiguity resolution. In Clifton, and Frazier (Eds), *Perspectives on sentence processing*. Hillsdale, NJ: Lawrence Erlbaum.

van Berkum, J, Brown, C, & Hagoort, P. (1999) Early referential context effects in sentence processing: Evidence from event-related brain potentials. *JML*, 41, 147-182.

Synfire Chains and Catastrophic Interference

Jacques P. Sougné and Robert M. French {J.Sougne, rfrench}@ulg.ac.be
Department of Psychology, University of Liège
4000 Liège, Belgium

Abstract

The brain must be capable of achieving extraordinarily precise sub-millisecond timing with imprecise neural hardware. We discuss how this might be possible using synfire chains (Abeles, 1991) and present a synfire chain learning algorithm for a sparsely-distributed network of spiking neurons (Sougné, 1999). Surprisingly, we show that this learning is not subject to catastrophic interference, a problem that plagues many standard connectionist networks. We show that the forgetting of synfire chains in this type of network closely resembles the classic forgetting pattern described by Barnes & Underwood (1959).

Introduction

A professional pitcher can send a baseball hurtling towards a batter at speeds approaching 100 miles an hour. In a mere 2 milliseconds the ball moves three inches, more than the width of a baseball bat. Given the long chain of neurons that must fire sequentially with incredible precision in order for the bat to connect with the ball, how could anyone ever hit a baseball? Consider gymnastics. The landing of a skilled gymnast dismounting from the high bar with a triple somersault is completely determined by the millisecond-precise instant he releases the bar. How is it possible to achieve the extraordinary timing accuracy necessary to consistently hit this beautiful landing correctly?

The problem is to find a way to make imprecise neurons act in an extremely precise manner. Nature has clearly found a way to circumvent the imprecision of individual neuron firings. The solution seems to rely on the presence of large *populations* of interacting neurons. In this paper we will discuss a mechanism for achieving precise timing, *synfire chains* (Abeles, 1991), that has received considerable empirical support. We will consider how the brain might learn these synfire chains and will present a neurobiologically plausibly computer simulation of synfire chain learning (see Sougné, 2001). Most importantly, we will show that the problem of catastrophic interference, a problem that plagues many types of neural networks (see French, 1999, for a review), does not seem to be a problem for synfire chains implemented in a sparsely-distributed network of spiking neurons. We simulate the classic forgetting experiment of Barnes & Underwood (1959) on a network designed to learn synfire chains and show that forgetting of information encoded in these simulated synfire chains very closely resembles the forgetting patterns of humans, as demonstrated by Barnes & Underwood. This leads to the prediction that real neural synfire chains will be forgotten gradually, rather than catastrophically.

Synfire Chains

Empirical data demonstrate the existence of very precise temporal behavior in neuron firings. For example, researchers have recorded spike timing of different cortical cells in monkeys (Abeles, 1991, Prut & al, 1998) and have observed the following stimulus-dependent pattern: when an initial neuron, A, fired, a second neuron, B, would fire 151ms later, followed by a third neuron, C, that would fire 289ms later *with a precision across trials of 1 ms*! Intervals of this duration require dozens of transmission delays from neuron A to neuron *C*. One of the major hypotheses about how this phenomenon could occur involves so-called "synfire chains." Abeles (1991). The other hypothesis is based on an increase in a population rate, which builds excitation in a downstream population, which, in turn, increases its firing rate, etc. (see Shadlen & Newsome, 1994). According to Abeles' hypothesis, since cortical synapses are relatively weak, many inputs to cells must arrive at the same time for them to fire. Consequently, each step in the synfire chain requires a significant pool of neurons whose simultaneous firing raises the potential of the next pool of neurons to allow them to fire. Recent experiments (Prut & al, 1998) indicate that these precise temporal firing sequences correlate more to behavior than to rate modulation and do not seem to be a byproduct of rate modulation. This would seem to buttress the synfire chain hypothesis.

Previous work on synfire chain learning has focused on how they can develop from a chaotic net with an unsupervised Hebbian learning rule (Bienenstock, 1995; Hertz & Prügel-Bennet, 1996). These studies involved an external stimulus which makes a large pool of neurons fire simultaneously at a particular instant. Subsequently, a sequence of successive large pools of simultaneous neuron firings occurs, produced by the random connection weights of the network. A given neuron will only fire if a large enough number of its presynaptic neurons provoke an increase of postsynaptic potential *simultaneously*. Connections are modified by a Hebbian learning rule. After learning, when the previously learned stimulus is presented again, the same chain fire, thereby constituting a synfire chain. These studies show that these chains are stable,

noise tolerant and that one network can store many different chains. Formal analysis showed that there is a relation between network size and the length of learnable synfire chains (Bienenstock, 1995), and that the recall speed should be faster than the training speed of the sequence (Sterratt, 1999).

In previous modeling work (Sougn□, 2001), it has been shown how a synfire chain can develop, thereby linking two pools of neuron firings caused by two sequential external stimuli. After learning, the first external stimulus will reactivate the stored synfire chain. It was also shown that synfire chain learning depends on the size of the network, the presence of long term depression (LTD) and the sparseness of connections. It turns out, surprisingly, that these synfire chains are not subject to catastrophic interference.

Catastrophic Interference

Gradual forgetting is one of the fundamental facts of cognition, which means that plausible models of human cognition must exhibit progressive forgetting of old information as new information is acquired. Only rarely does new learning in natural cognitive systems *completely* (or □catastrophically□) interfere with previously learned information (see, for example, French & Ferrara, 1999). However, it turns out that for a very large class of commonly used connectionist models □ those with a single set of shared (or partially shared) multiplicative weights (and most notably, standard feedforward backpropagation networks) □ learning new information can quite easily completely destroy all traces of previously learned information (McCloskey & Cohen, 1989; Ratcliff, 1990). In fact, the very features that give these connectionist models of memory their much-touted abilities to generalize, to function in the presence of degraded input, etc., are the root cause of catastrophic forgetting (See French, 1999, for a review of research on catastrophic interference).

Catastrophic interference is a radical manifestation of a more general problem for connectionist models of memory □ in fact, for *any* model of memory □, the so-called □stability-plasticity□ problem (Grossberg, 1982). The problem is how to design a system that is simultaneously sensitive to, but not radically disrupted by, new input.

A number of ways have been proposed to avoid the problem of catastrophic interference in connectionist networks. In the connectionist network that is our brain, McClelland, McNaughton & O□Reilly (1995) proposed that the dual memory system consisting of our hippocampus and neocortex evolved, at least in part, in order to overcome the problem of catastrophic interference (see also French, 1997).

In what follows, however, we hope to show that, rather unexpectedly, there is no catastrophic interference in the implemented network of old information from newly learned precise firing sequence. We will begin by considering a now classic experiment on human forgetting by Barnes & Underwood (1959). We will then show that when this experiment is simulated for synfire chains in a sparsely distributed network of spiking neurons, the forgetting curves observed during new learning are largely the same as those observed in Barnes & Underwood.

Forgetting Caused by New Learning

Barnes & Underwood (1959) conducted a series of experiments that measured the extent of retroactive interference in human learning. We will consider two of their experiments in this paper. In the first, subjects first learned a set of paired associates (A-B) consisting of a nonsense syllable and an adjective (e.g. *dax* paired with *regal*, etc.) and then were asked to learn a new set of paired associates (A-C) consisting of the same nonsense syllables associated with a new set of adjectives (e.g. *dax* paired with *dirty*, etc.). (This was called the A-B/A-C paradigm.) The forgetting curve for the A-B associate pairs produced by interference from the learning of the new A-C pairs was relatively gradual (Figure 1a).

In a second experiment, participants first learned a list of paired associates A-B, as above. Then they were asked to learn a series of paired associate where the second word was very semantically close to the original I word in the A-B pairs. They called this paradigm the A-B/A-B′ paradigm. See Figure 1b for the results of this second experiment.

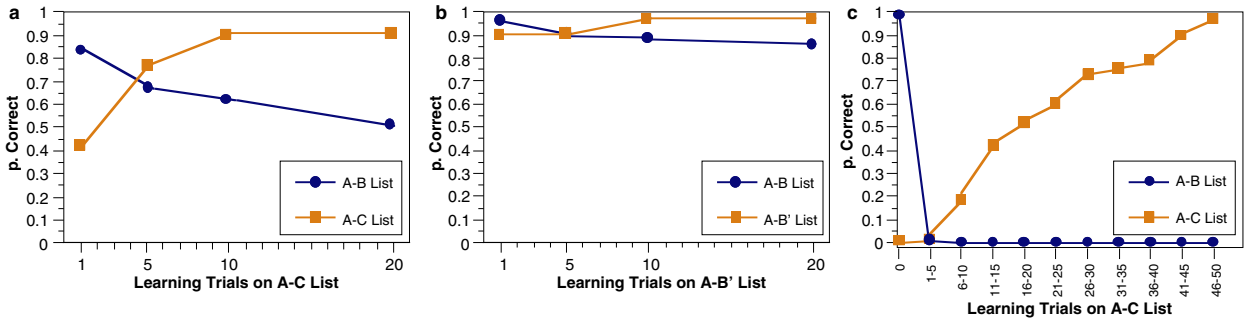


Figure 1, a: Gradual forgetting of previously learned information in Barnes & Underwood's (1959) A-B/A-C. b: Results of the same experiment when the C list closely resembles the original B list (A-B/A-B′ paradigm). c: McCloskey & Cohen's (1989) results showing the network's catastrophic forgetting A-B/A-C paradigm.

When connectionist networks began to become widely used as models of human memory, McCloskey & Cohen (1989) used the Barnes & Underwood A-B/A-C paradigm to test forgetting in these networks. It came as a considerable surprise to most researchers in the field that, at least under certain circumstances, McCloskey & Cohen were able to show that forgetting in a standard backpropagation network was anything but gradual. In one set of experiments, for example, a standard backpropagation network thoroughly learned a set of "one's addition facts" (i.e., the 17 sums 1+1 through 9+1 and 1+2 through 1+9). Then the network learned the 17 "two's addition facts" (i.e., 2+1 through 2+9 and 1+2 through 9+2). Recall performance on the originally learned one's facts plummeted as soon as the network began learning the new two's addition facts. Within 1-5 two's learning trials, the number of correct responses on the one's facts had dropped from 100% to 20%. In five more learning trials, the one's knowledge was at 1%, by 15 trials, no correct answers from the previous one's problems could be produced by the network. The network had "catastrophically" forgotten its one's sums. (See Figure 1c).

Networks of Spiking Neurons

In a network of spiking neurons (Maass & Bishop, 1999), nodes can be in two different states: they can fire (on), or they can be at rest (off). A node fires at a precise moment and transmits activation to other connected nodes with some time course. When a node activation or potential $V_i^{(t)}$ reaches a threshold, it emits a spike. After firing, the potential is reset to some resting value V_r. Inputs increase the node potential, but some part of the node potential is lost at each time step. Spiking neuron models, and in particular, INFERNET, the network discussed here, use a quite realistic post synaptic potential (PSP) function.

INFERNET is not a fully connected network; its structure is organized by clusters of nodes which constitute subnets. Each subnet is fully connected. From each node of a subnet there is a connection to every other node within that subnet. Some subnet nodes have connections to external subnet nodes. This not only reduces the computational demands of the program, but also better corresponds to the actual organization of the brain.

Two variables affect each connection: weight and delay. Each weight corresponds to the synaptic strength between a presynaptic and postsynaptic cell. The weight between a presynaptic node *j* and a postsynaptic node *i* is designated by w_{ij}. Noise is added to this value and the resulting noisy connection is denoted by $\hat{w}_{ij}$. The delay *d* of a connection determines when the effect of the presynaptic node firing will be maximum on the postsynaptic node. There is also a noise factor on the delay. This delay corresponds to the axonal, synaptic and dendritic delays of real neurons.

A signal, whether excitatory or inhibitory, will be affected by a leakage factor. When the signal has reached its maximum, at each following step of 1 ms, the signal will be divided by 2. Delays and leakage factors define the Post Synaptic Potential function $\varepsilon_{ij}(x)$:

$$\varepsilon_{ij}(x) = \frac{1}{2^x} H(x) \qquad (1)$$

$$\text{where: } H(x) = \begin{cases} 1 \text{ if } x \geq 0 \\ 0 \text{ if } x < 0 \end{cases}$$

and x is the difference between the time t, the time of the presynaptic node firing, and the noisy delay on the connection: $x = t - t_j^{(f)} - d$.

When a node potential V_i reaches a threshold θ, it emits a spike. Thereafter, the potential is reset to its resting value. After emitting a spike, a node enters a refractory period. This corresponds to the membrane resistance of real neurons which increases after a spike. In INFERNET, the refractory state of node *i* depends only on the last spike of the node *i*: $t_i^{(f)}$. A value dependent on the refractory state is subtracted from the node state value V_i. This value is denoted by $\eta_i(u)$, where u is the difference between the current time t and the time of the last spike of node *i*: $u = t - t_i^{(f)}$, and where a and b are constants

$$\eta_i(u) = \left[1 + e^{-\left(u^a - b\right)}\right] \vartheta(u)\theta \qquad (2)$$

$$\text{where: } \vartheta(u) = \begin{cases} +\infty, & \text{if } u < 1 \\ 1, & \text{otherwise} \end{cases}$$

Variables affecting the potential of a node have now been defined. Equation (3) express how $V_i^{(t)}$ is calculated at each time step.

$$V_i^{(t)} = \left[\sum_{j \in \Gamma_i} \sum_{t_j^{(f)} \in F_j} \hat{w}_{ij}\, \varepsilon_{ij}(x) \right] - \eta_i(u) \qquad (3)$$

Node *i* fires when its potential $V_i(t)$ reaches the threshold Θ. This potential is affected by connection weights $\hat{w}_{ij}$ coming from each presynaptic node *j*. The set of presynaptic connections to node *i* is given by $\Gamma_i = \{j | j \text{ is presynaptic to } i\}$. F_j is the set of all firing times of presynaptic nodes *j*: $t_j^{(f)}$. Noisy connection weights linking *j* node to *i* node are $\hat{w}_{ij}$.

Learning

Long term potentiation (LTP) and depression (LTD) are the basic mechanisms of long-lasting modifications of synaptic efficiency. Hebb (1949) postulated that when presynaptic activity coincides with postsynaptic activity, the connection between both neurons is strengthened. According to recent experiments, the modification of synaptic efficiency depends on precise timing of afferent signals (neurotransmitters binding to

receptors) and the postsynaptic neuron spike. LTP seems to require that postsynaptic action potential be simultaneous or subsequent to postsynaptic currents (Markram et al., 1997; Zhang et al., 1998). In short, when the signal from the presynaptic neuron firing arrives before, or during the spike of postsynaptic neuron, the synapse is strengthened (LTP). When the signal from the presynaptic neuron arrives after the spike of postsynaptic neuron, the synapse is depressed (LTD).

In the present model, the plasticity of a synapse w_{ij} is a function of three parameters: the firing time of the presynaptic neuron: $t_j^{(f)}$, the transmission delay between this firing and its effect on the postsynaptic neuron (d_{ij}), and the firing time of postsynaptic neuron $t_i^{(f)}$. Learning in INFERNET consists of modifying the weights of connections between nodes w_{ij} by a value $\check{}\, w_{ij}$ (weights are all short integers from -32767 to 32767, which explains why the weight change values run from –1024 to 1024). The Hebbian learning function used is shown in Figure 2. This function follows empirical studies (Markram et al., 1997; Zhang et al., 1998). Similar functions have been used in various simulations by others (Levy & Horn, 1999; Munro & Hernandez, 1999).

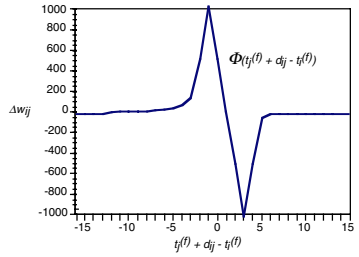


Figure 2: INFERNET's Hebbian learning function: when the signal from the presynaptic neuron arrives before or during the spike of the postsynaptic neuron, the synapse is strengthened (LTP); when the signal arrives after this spike, the synapse is depressed (LTD).

The learning algorithm attempts to reproduce the temporal relation between two successive inputs. This is particularly difficult because two successive inputs can be separated by several tenths of a second and a single connection cannot alone be responsible for such long delays. A long chain of successive pools of node firings is therefore required. This problem is illustrated in Figure 3. The problem is linking nodes *a* and *a´* that fire at time 0 with node *g* firing at time 49. In the learning phase, only nodes *a* and *a´* and *g*, 49 ms later, are externally stimulated. The system has to find a chain of node-firings that makes the target node *g* fire at time 49 when the probe nodes *a* and *a´* fire at time 0. The levels shown in Figure 3 are defined by the pools of firing nodes that separate the nodes firing in response to the input probe and the nodes responding to the target. Note that when simultaneous input from enough afferent neurons does not occur, the node will not fire. There is therefore a phenomenon of selection of only those nodes that have fired due to simultaneous inputs. This requires a large fan out of connections at all levels between the probe nodes and the target nodes.

The refractory state indicates when a particular node fired. We also know the delay of signal propagation from a presynaptic node to a postsynaptic node. From these two values we can, therefore, detect which synapse can contribute to a node firing at the right moment. In Figure 3, one can detect which nodes contribute to the firing of node *g* i.e., *e* and *f*, whose signals arrive at *g* virtually simultaneously. If *f* fired 9 ms before *g*, and *e* fired 11 ms. before *g*, their respective connections will be strengthened. Similarly, one can also determine which nodes contributed to the firing of *f* (i.e., *d* and *d´*), if *d* fired 7 ms before *f*, their connection will be strengthened (to a somewhat lesser extent because we are farther down the chain). This chaining rule acts as if a signal was going backwards from the target node to the probe nodes, losing a bit of its strength at each step. In order to reduce combinatorial explosion, only the *n* best contributing nodes are selected for the next level in this chaining rule. Connections between nodes will be modified according to equation (4):

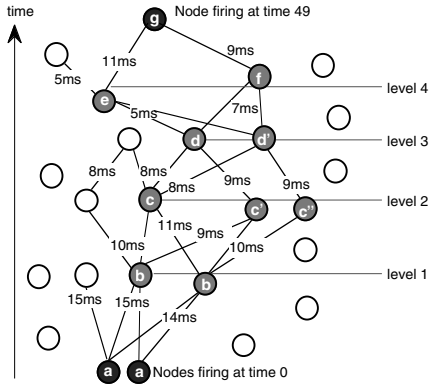


Figure 3: The chaining rule problem: Learning a path of neural firings that makes node *g* fire exactly 49 ms after nodes *a* and *a´*.

$$\Delta w_{ij} = \Phi\left(t_j^{(f)} + d_{ij} - t_i^{(f)}\right) - \lambda \qquad (4)$$

$$where \quad \lambda = \left\{ \begin{array}{c} -level \ if \ \Phi\left(t_j^{(f)} + d_{ij} - t_i^{(f)}\right) is \ negative \\ level \ otherwise \end{array} \right\}$$

This rule is based on the history of node firing and has neurobiological justification. For example, Markram, et al. (1998) show that the state of a synapse is indicative of its past activity. Moreover, empirical studies (Engert & Bonhoeffer, 1997) show that LTP also propagates from the originating synapse to neighboring synapses, lending further plausibility to the present chaining rule. In addition, each connection has a small decay factor (of –10 by epoch).

The learning algorithm is triggered only when external input is presented. We can imagine that

external input provides a strong signal that triggers the chaining rule. Note that Hebbian learning does not seem to be dependent on this kind of signal and affects probably all synapses downstream from an action potential. Here, it is the target input that is the signal to launch the chaining rule. The objective is to link the probe nodes firing to the target nodes firing and to avoid reinforcing other irrelevant firings.

Simulation 1

In general, motor forgetting occurs more slowly than cognitive forgetting (Globerson, Nahumi, Ellis, 1998). By testing the present synfire chain algorithm for cognitive forgetting,, we reasoned that if catastrophic interference disappeared for this paradigm, the same algorithm would eliminate it for precise motor learning.

The following simulation is based on the original AB-AC paradigm used in Barnes & Underwood (1959). As in the original experiment, we created a list of non-words (A) and two associated lists of words (B and C). Each B and C word was coded over 6 nodes (out of 800 possible nodes) and each non-word A was coded over 16 nodes. Although the selection of nodes was made randomly, we ensured that there was very little overlap among the items in the A-list and among the items in the B-list. All lists consisted of eight items.

Each temporal firing sequence consisted of 16 nodes firing at time 0, corresponding to a presentation of a non-word from list A in the Barnes & Underwood experiments. Six nodes fired at time 60, triggered by the associated word from the B wordlist.

Once the network had learned to associate the items in list A with those in list B, the network then had to associate the node-firings associated with the items in the A-list with those of the C-list. As in Barnes & Underwood, we kept the similarity very low between the corresponding words in the B and C lists. This meant that very few nodes overlapped in the encoding of the corresponding words from the two lists. As the network learned the new set of associations, we tracked how fast new learning was taking place (i.e., how close the output of the network was to the desired word in List C) and, at the same, time, how far the output was from the originally learned word in List B.

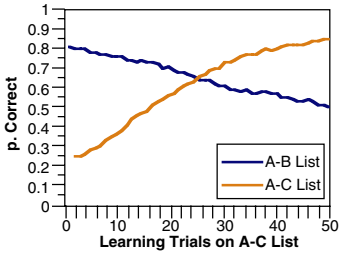


Figure 4: INFERNET performing AB-AC learning.

The results, based on 20 runs of the program, are shown in Figure 4. The Y axis indicates the proportion of correct node firings, i.e. the number of B and C node-firings within a –2ms window divided by the total number of nodes in B and C words. It is clear that, unlike the catastrophic interference observed in standard backpropagation networks (see Figure 1c), this sparsely-distributed network of spiking neurons can learn the second set of associations without catastrophically forgetting the previously learned list. These results are strikingly similar to those of Barnes & Underwood (Figure 1a).

Simulation 2

All of the model parameters for this simulation were identical to those of the preceding simulation, with the exception that wordlist C was replaced by a wordlist B′. All words in the B′ List were very similar to the corresponding words in the B list. Of the 6 nodes used by the representation of each B′ word, 4 of them were shared by the corresponding word in the B List.

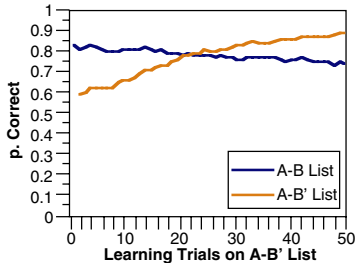


Figure 5: INFERNET performing AB-AB′ learning.

The results, based on 20 runs of the program, are shown in Figure 5. Again, this simulation closely reproduced the experimental results of Barnes & Underwood's second experiment (see Figure 1b). The results indicate the second associations are learned more quickly and the first associations are almost not forgotten, as for humans.

Conclusions

Human learning involves relating two signals separated in time, or linking a signal, an action and a subsequent effect. On occasion, the precise timing of these signals is of critical importance. A millisecond inaccuracy can mean that the spear thrown by the hunter will miss its target, that the gymnast will miss her landing, etc. Events may often be separated in time, but nonetheless, humans can link them, if necessary, with extraordinary accuracy, thereby allowing them to correctly perform a particular action at precisely the right moment. We have explored one major hypotheses concerning how the brain might achieve this namely, synfire chains.

Clearly people are not born with encodings of this timing information. Hunters *learn* to throw projectiles accurately, gymnasts *learn* to land correctly. Precise temporal firing sequences must be learnable and permit the linking of two events with extreme precision.

A learning algorithm based on a Hebbian learning rule has been presented in this paper. We have briefly explored the ability of a sparsely-distributed network of spiking neurons, INFERNET, to learn synfire chains and, most importantly, we studied forgetting in this network of these chains. Unlike many current connectionist networks, we found that the forgetting of synfire chains is not subject to catastrophic interference, but rather, closely resembles the gradual forgetting curves exhibited in Barnes & Underwood's (1959) paper on human forgetting. This is due to the sparseness of the number of paths (compared to the very large number of possible paths) from probe to target created by the learning algorithm. We hope to have demonstrated the importance of synfire chains for human cognition and to have shown an implementation in a network of spiking neurons. Finally, and crucially, our simulations indicate that synfire chains, so necessary for precision actions in the real world, may not be affected by catastrophic forgetting.

Acknowledgements

This research was supported by the Belgian PAI Grant p4/19 and the E.C. grant HPRN-CT-1999-00065.

References

Abeles, M. (1991). *Corticonics: Neural circuits of the cerebral cortex.* New-York: Cambridge University Press.

Barnes, J., & Underwood, B. (1959). "Fate" of first-learned associations in transfer theory. *Journal of Experimental Psychology, 58*, 97-105.

Bienenstock, E. (1995). A model of neocortex. *Network: Computation in Neural Systems, 6,* 179-224.

Engert, F., & Bonhoeffer, T. (1997). Synapse specificity of long-term potentiation breaks own at short distances. *Nature, 388,* 279-284.

French, R. M. (1997). Pseudo-recurrent connectionist networks: An approach to the "sensitivity-stability" dilemma. *Connection Science, 9,* 353-379.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences, 3,* 128-135.

French, R. M., & Ferrara, A. (1999). Modeling time perception in rats: Evidence for catastrophic interference in animal learning. In *Proceedings of the 21st Annual Conference of the Cognitive Science Conference*. NJ:LEA, 173-178.

Globerson, S., Nahumi, A., & Ellis, S. (1998). Rate of forgetting for motor and cognitive tasks. *International Journal of Cognitive Ergonomics, 2*, 181-191.

Grossberg, S. (1982). *Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition, and Motor Control.* Boston: Reidel.

Hebb, D. O. (1949). *The Organization of Behavior.* New York: Wiley.

Hertz, J.A., & Prügel-Bennet, A. (1996). Learning synfire chains by self organazation. *Network: Computation in Neural Systems, 7,* 357-363.

Levy, N., & Horn, D. (1999). Distributed synchrony in a Hebbian cell assembly of spiking neurons. In *Advances in Neural Information Processing Systems 11,* Cambridge Ma, MIT Press.

Maass, W., & Bishop, C. M. (1999). *Pulsed Neural Networks.* Cambridge, Ma, MIT Press.

Markram, H., Gupta, A., Uziel, A., Wang, Y., & Tsodyks, M. (1998). Information processing with frequency-dependent synaptic connections. *Neurobiology of Learning and Memory, 70,* 101-112.

Markram, H., Lübke, J., Frotscher, M., & Sakmann, B. (1997). Regulation of Synaptic Efficacy by coincidence of Postsynaptic Aps and EPSPs. *Science, 275,* 213-215.

McClelland, J., McNaughton, B., & O'Reilly, R. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102,* 419-457.

McCloskey, M., & Cohen, N. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (ed.) *The Psychology of Learning and Motivation: Vol. 24*, pp. 109-164, NY: Academic Press

Munro, P., & Hernandez, G. (1999). LTD facilitates learning in a noisy environment. In *Advances in Neural Information Processing Systems 11,* Cambridge Ma, MIT Press.

Prut, Y., Vaadia, E., Bergman, H., Haalman, I., Slovin, H., & Abeles, M. (1998). Spatiotemporal structure of cortical activity: Properties and behavioral relevance. *Journal of Neurophysiology, 79,* 2857-2874.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psych. Review, 97*, 285-308.

Shadlen, M. N., & Newsome, W. T. (1994). Noise, neural codes and cortical organization. *Current Opinion in Neurobiology, 4,* 569-579.

Sougné, J. P. (1999). *INFERNET: A neurocomputational model of binding and inference.* Doctoral dissertation, Université de Liège.

Sougné, J. P. (2001). A learning algorithm for synfire chains. In R. M. French, & J. P. Sougné (Eds.) *Connectionist Models of Learning, Development and Evolution.* pp. 23-32 London: Springer Verlag.

Sterratt, D. C. (1999). Is biological temporal learning rule compatible with learning synfire chains? *Proceedings of the Ninth International Conference on Artificial Neural Networks.*

Zhang, L. I., Tao, H. W., Holt, C. E., Harris, W., & Poo, M. (1998). A critical window for cooperation and competition among developing retinotectal synapses. *Nature, 395,* 37-44.

Human Sequence Learning: Can Associations Explain Everything?

Rainer Spiegel (RS272@CAM.AC.UK)
University of Cambridge, Department of Experimental Psychology,
Downing Street, Cambridge, CB2 3EB, UK

IPL McLaren (IPLM2@CUS.CAM.AC.UK)
University of Cambridge, Department of Experimental Psychology,
Downing Street, Cambridge, CB2 3EB, UK

Abstract

It will be shown that whilst a popular connectionist model, the simple recurrent network (SRN) as introduced by Elman (1990), is a very good first approximation in modeling human sequence learning, it is not, in itself, sufficient. At CogSci 2000, all five papers referring to the SRN tried to provide evidence that it is an adequate model of human performance. We will take on a more moderate position. The results of a human experiment followed by a structured interview reveal that human sequence learning is not always the kind of statistical process captured by the SRN alone.

Introduction

In cognitive science, there is an ongoing debate whether human learning should be modeled by the explicit use of rules or by figuring out statistical regularities. In addition, it has been emphasized that human learning consists of both rule and associatively-based processes (e.g. Jones & McLaren, 1999; McLaren et al. 1994).

Perhaps the most popular connectionist model used to study sequence learning is the previously mentioned SRN developed by Elman, which has become ubiquitous in the literature (Cleeremans, 1993; Elman, 1990; Elman et al. 1996; McLeod et al. 1998). A diagram of the SRN is shown in Figure 1.

The SRN is capable of learning any sequence, even sentences with hierarchical structure (McLeod et al. 1998). However, as will be seen later, this is not to be confused with learning to master any kind of sequential problem. Among other connectionist models, the SRN does not implement rules and therefore learns sequences in an associative way. Therefore, the results of the SRN should resemble human learning if humans also learn sequences associatively. In contrast, the results of the SRN might differ from human performance either if human sequence learning incorporates something more than an associative process, or if the associative mechanisms used in human sequence learning are not those employed in the SRN.

In the SRN, the network receives input from the input units and is made to predict the next step of the sequence at the output level. The SRN has connections from the hidden units to the so-called context units, which are exact copies of the hidden units one time step ago. All other connections in the network are adjustable. The context units provide the SRN with a dynamic memory, i.e. depending on the sequence position, the very same inputs can result in different predictions of the network. Each time step, the network is trained by adjusting the weights on the connections according to the backpropagation learning algorithm (first introduced by Werbos, Le Cun, Amari, Parker and now most widely accessible in Rumelhart et al. 1986). The SRN with a supervised learning algorithm was chosen for this paper, because it was considered appropriate in modeling the human experiment (as will be seen later, the subjects in the human experiment received a signal if they had made an error and no signal if they had made the correct response).

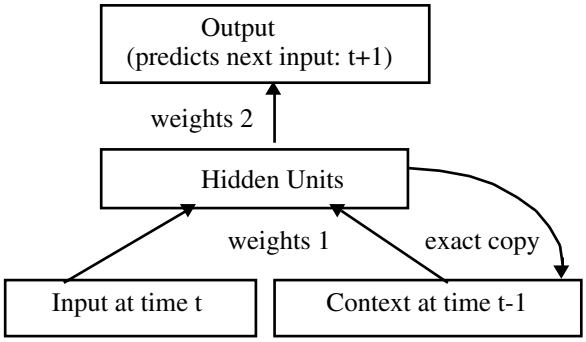


Figure 1: The simple recurrent network (SRN).

The serial reaction time task (Nissen & Bullemer, 1987) is a particularly successful paradigm used to test human sequence learning. In this type of task, the subject sits in front of a screen on which one of a number of lights flashes on at different locations. The subject is asked to press the key below the flashing light as fast as possible and the reaction time of each response is measured. While the subjects are not informed about any sequence in the stimulus material, the lights flash in a particular order. Therefore, there is a contingency in the way that the preceding stimuli

predict the current one. If subjects speed up their reaction times on the sequences but are not able to provide verbal information concerning the contingencies, then their learning could be considered associative. If subjects can verbalize the contingencies, their behavior can be seen as cognitive and/or associative, depending on the degree to which subjects are able to state the underlying rules governing the sequences. A particular advantage of the task is that the stimuli of the sequence are presented one after another and as a result, the reaction times to each separate stimulus can be measured. Consequently, it can be compared to each separate output activation of the SRN used to model the task. Another advantage is that there are lights flashing on at different screen locations rather than symbols. Symbols would have a semantic meaning for subjects, which would have the consequence that subjects would probably not start the experiment without preconceptions.

Simulation studies with the SRN

The SRN has been assumed to be able to learn any kind of sequence learning task with the exception of the catastrophic interference problem, i.e. when the SRN learns a set of sequences and is then trained on a new set of sequences similar to the old ones and tested again on the old sequences, its performance is very poor. The reason for this problem, however, seems to lie in the fact that the SRN uses backpropagation as the learning algorithm which does not have an adaptive learning rate. However, McLaren (1993) has shown that an adaptively parametrised error correcting system (APECS) can avoid catastophic interference. As a result, there seems to be some chance that all sequential problems could possibly be modeled successfully with neural networks employing the SRN architecture. Although there is published work about weaknesses of the SRN (Servan-Schreiber et al., 1991; Cleeremans, 1993; Maskara & Noetzel, 1993; Timmermans & Cleeremans, 2000), those and other reported failures did not turn out to be due to the SRN architecture *per se* (Spiegel, Jones & McLaren, 2001; Spiegel & McLaren, 2001). Our new papers may lead some connectionists to argue that the SRN models human performance entirely. In order to prevent this seemingly wrong conclusion from happening, we present a sequential problem that can be solved by humans, but not by the SRN. Furthermore, a detailed network analysis about how the SRN tackles the task will be necessary to prove that this problem cannot be completely solved by the SRN. We start with an analysis of SRN performance on the task that we eventually settled on.

Procedure

The SRN was implemented using the C programming language. The task can be represented by the following grammar: (ab(c*1 ^ 3)ba) ^ (abb(c*1 ^ 3)bba). Here, the symbol ^ stands for the word *or*. For connectionist models, those letters have no semantic meaning. Expressed in words, those grammars mean: The sequence starts with the letter A. After that, the letter B follows either once or twice. Then, the letter C follows either once or three times, before the letter B appears the same number of times it had appeared earlier (once or twice) and finally, the sequence ends with the letter A. The input and output layers have a local representation for each symbol in the sequence, i.e. a = (1,0,0), b = (0,1,0) and c = (0,0,1). The network is trained with a learning rate of 0.1 and 300 hidden units.

Results

As it turned out, the network could learn this problem, a not inconsiderable achievement as there had been claims that it could not (for an overview, see Spiegel et al. 2001), but never generalized to both novel sequences with the same structure, but two c's, i.e. (abccb**a**) ^ (abbccb**b**a). It failed to predict an a after the final b in the first sequence type and failed to predict the second b after the second last b in the second sequence type (the bold letters). An entire simulation experiment with eight networks having been trained for at least half a million trials never resulted in the case where the SRN generalized to both novel sequences, even when the lowest stringent criterion was set, i.e. best match.

Modifying network parameters After this simulation experiment, 300 other SRNs were run with different numbers of hidden units (ranging from 5 to 500) on the same problem, but none of them reached sufficient generalization performance. Moreover, with less than 6 and more than 450 hidden units, the network completely lost the ability to learn the task.

Network analysis A separate network analysis focused on the ability to generalize to varying numbers of c fillers in the range between two and eleven. An interesting discovery was made: the SRN would never stabilize in generalizing to any even number of c fillers in both sequence types of the grammar displayed earlier, but it would generalize after an odd number of c fillers. In essence, the network appeared to exploit the fact that it was only trained on an odd number of c filler items by adopting cyclical patterns of activity tuned to the 1 and 3 c cases. These would also apply to other odd numbers of c items (e.g. 101), but never to even numbers of the c fillers. Moreover, whilst the performance on the trained patterns would remain stable, the ability to generalize to novel patterns would fluctuate over training trials.

The possibility of generalization during transitions between these stable states remains, however, so further tests were carried out. To assess this possibility, a very sensitive network session had to be run, with different

numbers of hidden units and a test of generalization performance after each single trial. As a consequence, an SRN was implemented that would do 100,000 generalization tests during 100,000 training trials. Finally, two cases were found: after trial 39956 a 400 hidden units SRN fulfilled the best match criterion on generalizing to both two c sequences while also mastering the sequences it had been trained on. One trial later, it had lost its ability to generalize. On trial 39967 it regained this ability for one single trial, but lost it immediately from the next trial onwards and never regained it. Hence, we thought this was enough evidence to state that the SRN does not stabilize in generalizing to the sequences with two c's. Based on those findings, an experiment with human subjects was carried out to explore whether they were able to generalize to the two c case.

Human Experiment

The experiment comprised a three choice serial reaction time task. The stimulus was a circle flashing in different locations on a computer screen. The circles were arranged as a triangle, i.e. lower left corner, upper middle corner, lower right corner. The subjects were asked to press the key that corresponded to the stimulus location as fast and as accurately as possible. They were divided into an Experimental and a Control group. In both groups the order of presentation during training blocks as well as during testing followed a sequence. The sequences for the Experimental group were the same as those that the SRN had been trained on. They shall be called *consistent* sequences from now on:

(ab (c*1 ^ 3) b**a**) ^ (abb (c*1 ^ 3) b**b**a)

In the human experiment, all three letters corresponded to a particular circle, i.e. circle flashes were what the subjects saw, not letters. In the first sequence type, subjects should learn to predict the final a (bold letter) once the c had stopped and the letter b had appeared. In the second sequence type, subjects should be able to predict the second b (bold letter) once the letter c had stopped and the first b had appeared. The Control group received the same sequences as the Experimental group in 50 percent of the cases, and the following ones in the other 50 percent of the cases. They shall be called *inconsistent* sequences from now on:

(ab (c*1 ^ 3) b**b**) ^ (abb (c*1 ^ 3) b**a**a)

Because the final letter in the first sequence type and the letter before the final letter in the second sequence type had an alternative letter in 50 percent of the cases, the Control group should never be able to predict the location of the last circle in the single b case and the location of the circle before the last circle in the double b case. There were four training sessions of equal length for both Experimental and Control groups. Following that, there were two testing sessions of equal length in which both groups received 50 percent of the following consistent sequences:

(ab (c*1 ^ 2 ^ 3) b**a**) ^ (abb (c*1 ^ 2 ^ 3) b**b**a)

In addition, both groups received 50 percent of the following inconsistent sequences:

(ab (c*1 ^ 2 ^ 3) b**b**) ^ (abb (c*1 ^ 2 ^ 3) b**a**a)

The difference between the training trials and the testing trials lies in the fact that both groups receive the same sequences during testing and both groups receive the two c case which is used to test their performance to generalize to novel sequences.

The experiment aimed to investigate the following hypotheses: subjects in the Experimental group should perform faster on the critical positions (=bold letters) in the consistent sequences than in the inconsistent sequences and they should generalize to the novel sequences, because they were constructed according to the same underlying grammar. On the other hand, subjects in the Control group should show no real difference between the reaction times on consistent and inconsistent sequences. The same holds for accuracy. Subjects in the Experimental group should be more accurate on the critical positions of the consistent sequences, whereas subjects in the Control group should show more or less equal accuracy on consistent and inconsistent sequences. As a result, the (RTinconsistent-RTconsistent) differences as well as the (Errors_inconsistent-Errors_consistent) differences should be significantly higher in the Experimental group than in the Control group.

Method

Subjects The experiment comprised 30 subjects aged 18 to 40 years who were either graduate or undergraduate students at the University of Cambridge. The subjects were randomly assigned to each condition.

Apparatus The experiment was run on a Macintosh Quadra 630 computer. The subjects were seated approximately 80cm from the screen, which was roughly at eye level. The diagonal of the screen was 30cm in size. The light in the room was dimmed to a constant level.

Procedure After detailed instructions, the circles appeared on the screen. The display consisted of white outlines of three triangularly placed circles in the middle of a black background. They were two centimeters in diameter and the centers of the circles on the bottom of the triangle were approximately 5.5cm

apart. The center of the upper circle was approximately 4cm apart from the centers of the two other circles. Each trial, one of the outlines would flash in such a way that it would become a solid white circle that remained on the screen until the subject responded or had not pressed a key within 4.25 seconds of the stimulus onset. After each response or after 4.25 seconds the solid circle was immediately cleared leaving only the outlines remaining. The response keys were arranged in the following way: the lower left circle corresponded to the 'v' key, the upper middle circle to the 'b' key and the lower right circle to the 'n' key. Subjects were requested to use their index-, middle-, and right finger of their preferred hand. If subjects took longer than 4.25 seconds, pressed the wrong key or a different key than the three designated, an acoustic signal indicated that the subject had made an error. Reaction time was measured in milliseconds from the stimulus onset until the key press, and the interval between a response and the onset of the next stimulus was 180ms. When one sequence finished, the screen was cleared (to the black background) and then the three outlines reappeared for 600ms until the first circle filled with white.

Block characteristics Both Experimental and Control group started with one block of 9 random circle locations in order to assess the subjects' baseline reaction time and accuracy. Then both Experimental group and Control group received 78 sequences in each of the six following blocks. Out of those six blocks, the first four blocks comprised the training trials and the last two blocks comprised the testing trials. Out of the 78 sequences in both training and testing phase, the first six of each block were not taken into the final analysis because concentration at the beginning of each block may be influenced by the preceding pause. Of the remaining 72 sequences in each block of the training phase, the Experimental group randomly received eighteen of all possible combinations of the consistent sequences (see above). The Control group randomly received nine of all possible combinations of both consistent and inconsistent sequences.

Of the 72 sequences in the testing phase, both Experimental and Control group received six of all possible combinations of the earlier mentioned consistent and inconsistent sequences per block, i.e. 12 as a whole (in order to determine the average reaction time, only the reaction times where the subjects made a correct response were counted):

The difference between these trials and those of the Control group in the training phase was that it tested whether the subjects were able to generalize to novel sequences, i.e. the ones containing the letter c twice.

Interview A structured questionnaire immediately followed the last block of the experiment. In this questionnaire, subjects were asked whether the circles had flashed on in a particular sequential order, and if so, what they can tell about the sequences.

Results

It was necessary to assess both average reaction time and number of error differences because a significant result for one of the measures does not necessarily mean much, as there could be a significant opposite trend in the other. If this was the case, the significant result in the expected direction would not reveal evidence for learning. This effect is called speed-accuracy tradeoff.

Reaction times The results of the average reaction times are considered first, i.e. the dependent variable was: inconsistent minus consistent average reaction time. An analysis of variance was carried out with the between subjects factor *group* (Experimental vs. Control group) and the within subjects factors *type* (single vs. double b case) and *number of c's* (one vs. two vs. three). Before this analysis was carried out, we tested whether the underlying assumptions for an analysis of variance with repeated measurements were met. Cochran's C test to check the equality of variances as well as the Mauchly test of sphericity (Norusis, 1994) revealed that the assumptions were entirely fulfilled. The analysis of variance revealed a significant main effect for the between subjects factor *group*, $F(1,28)=9.35$, $p<.01$, $f=.57$. The Experimental group ($M_e=34.71$, $\pm SE_e=7.58$) reveals a significantly higher reaction time difference when compared with the Control group ($M_c=1.49$, $\pm SE_c=7.79$). The size of this effect is expressed in terms of the index f (Cohen, 1988). Because f-values greater than .4 are considered large, this effect can be regarded as very strong. Cohen's f can be set in relation to more traditional effect size measures, such as the amount of variance explained by this effect ($\eta^2=.25$).

Thereafter, it was crucial to know whether subjects show a reliable effect on the one c case and the three c case and in particular whether they are able to generalize to the novel sequences with two Cs. This last will form the basis for the critical comparison with the SRN. In this experiment, the two c case showed a result on the borderline between significant and marginally significant, $F(1,28)=2.82$, $p=.05$, ($M_{e\_2c}=21.65$, $\pm SE_{e\_2c}=9.22$ vs. $M_{c\_2c}=-4.0$, $\pm SE_{c\_2c}=12.14$), providing some evidence that people do generalize to novel sequences. It is interesting to note that the sequences the subjects had been trained on obviously show larger effects, which is partly reflected in the strong main effect of the ANOVA. In order to provide a better comparison between trained and novel sequences, here are the results for the sequences the subjects were trained on:

The one c case revealed a significant effect in favor of the Experimental Group, $F(1,28)=5.29$, $p=.01$,

(M_{e_1c}=47.23, ±SE_{e_1c}=15.77 vs. M_{c_1c}=-4.75, ±SE_{c_1c}=16.24). Similarly, the three c case showed a significant effect in the same direction $F(1,28)=4.62$, $p=.02$, (M_{e_3c}=35.26, ±SE_{e_3c}=7.84 vs. M_{c_3c}=13.22, ±SE_{c_3c}=6.61). The full results are displayed in Figure 2.

Number of Errors So far, however, it could still be that the effects on the reaction times are due to a speed-accuracy tradeoff. Therefore, it was necessary to focus on the second performance measurement, i.e. the *number of errors* subjects made with consistent and inconsistent sequences. The same kind of ANOVA with the dependent variable *error differences* revealed no significant difference between Experimental and Control group, $F(1,28)=2.28$, ns., nor did any of the individual comparisons for all three numbers of c even show a descriptive trend in the opposite direction, which entirely excludes the possibility of a speed-accuracy tradeoff.

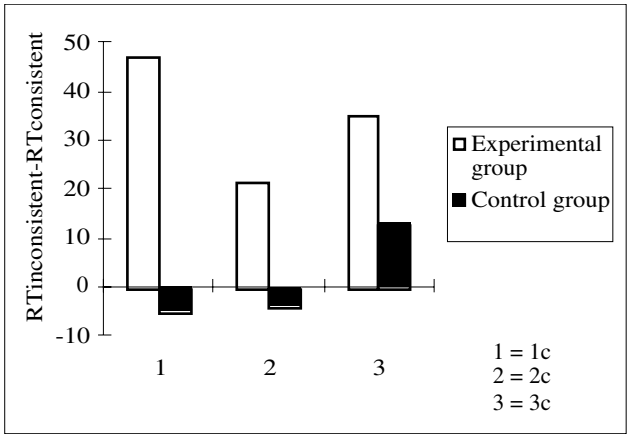


Figure 2: Average reaction time differences in humans.

Computational Model The Experimental SRN was trained with the same sequences as the Experimental group until it first reached the earlier defined performance criteria. The Control SRN was trained for 40,000 trials on the same sequences as the Control group. Both SRN's were trained with a learning rate of .1 and had 300 hidden units. The results of the network performance are displayed in Figure 3.

As can be seen in terms of the output activation differences (activation corresponding to the critical target value of the consistent sequences minus activation corresponding to the critical target value of the inconsistent sequences), the Experimental SRN has learned the task, but is not able to generalize to the two c case in any way. The Control SRN more or less resembles the human Control group in a way that it is not able to predict the next position, because it equally favors consistent and inconsistent sequences.

Structured Interview In order to get a better idea of how people solved this task, it was necessary to find out to what extent subjects verbalize the sequences. Here it was crucial to find out how many people in the Experimental group verbalized the rule that the number of b's after the c's was dependent on the number of b's before the c's. Only fourteen out of fifteen subjects in the Experimental group were able to take part in the interview. All fifteen subjects in the Control group answered the questions.

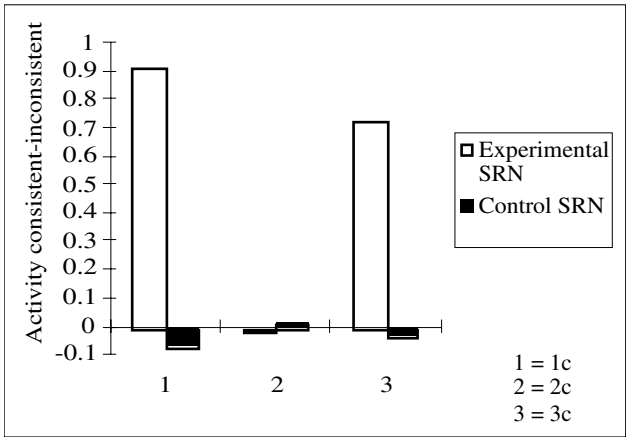


Figure 3: Activity differences between consistent and inconsistent output units on the critical letters.

Almost all of the subjects in the Experimental as well as the Control group verbalized that the circles flashed in a sequential order, while none of the subjects in either group was able to verbalize how many times a circle flash corresponding to the grammatical letter C had appeared.

None of the subjects in the Control group verbalized any dependency between the b flashes before and after the c flashes, which was expected, because they were independent of each other in the Control group. In the Experimental group, two out of fourteen subjects verbalized this dependency. A close look at the reaction time differences of those two subjects revealed that their reaction time differences were more pronounced in their effect (1c=107.32, 2c=51.37, 3c=82.55) than the average RT-differences of the remaining twelve subjects in the Experimental group (1c=32.56, 2c=16.72, 3c=25.12) and none of the error differences point in the wrong direction, which excludes the possibility of a speed-accuracy tradeoff. Interestingly, when performing the same ANOVA with the exception of the two subjects who verbalized the rule, there is still evidence of learning the trained sequences: $F(1,25)=4.71$, $p<.05$, $f=.43$, $\eta^2=.16$, but no longer evidence for generalization to the novel sequences, of $F(1,25)=1.54$, ns., (M_{e_2c}=16.71, ±SE_{e_2c}=10.91 vs. M_{c_2c}=-4.0, ±SE_{c_2c}=12.14). In other words: Successful learning but generalization failure occurs when the subjects who represented the rule are left out of the analysis, which corresponds to the results of the

associative SRN. As a result, one tentative conclusion is possible here: the human ability to represent the rule (which is absent in the SRN) may have led to successful generalization in humans. There is, of course, another equally valid interpretation of this finding, however, and that is that those subjects who had learned the sequences (associatively) best were the ones who subsequently became aware of them and were able to induce the rules governing them.

Discussion

This experiment provides evidence that humans and the SRN may differ when dealing with a particular sequential problem. Whilst the SRN is capable of learning all of the sequences presented in the training set, it cannot generalize to particular sequences that were constructed according to the same underlying grammar. Furthermore, a logical analysis of the inner representations of the network revealed the reason why it does not learn the problem: the way the network represents the temporal order of the sequences in its context layer makes it impossible to solve the complete generalization problem.

There is some evidence that humans approach the problem in a different way. The structured interview suggested that some humans can induce the underlying rules of the sequences and the results of those subjects in the experiment provide evidence that they may make explicit use of them when generalizing to novel sequences.

However, it would be hard, and possibly premature to uncouple the rule-based and the associative component of humans who have participated in this task. Those two subjects who represented the task in a rule-based way have probably started with associative learning and later somehow induced the rule. The suggestion here is that there is a real possibility that there are associative mechanisms available to humans which interact with cognitive processing to determine task performance. On the basis of our results, we consider the purely associative SRN a very powerful model that may be able to learn many kinds of sequence, but does *not* simulate the human ability to generalize in this experiment.

Acknowledgments

We gratefully acknowledge the support from Emmanuel College Cambridge and the Cambridge European Trust.

References

Cleeremans, A. (1993) *Mechanisms of Implicit Learning. Connectionist Models of Sequence Processing*. Cam- bridge, MA: MIT-Press.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, N.J: Erlbaum.

Elman, J.L. (1990) Finding structure in time. *Cognitive Science, 14*, 179-211.

Elman, J.L., Bates, E.A., Johnson, M.H., Karmiloff-Smith, A., Parisi, D. & Plunkett, K. (1996*) Rethinking Innateness: A connectionist perspective on development*. Cambridge, MA: MIT-Press.

Jones, F.W. & McLaren, I.P.L. (1999) Rules and Associations. *Proceedings of the Twenty-First Annual Con- ference of the Cognitive Science Society, pp. 240-245.* Mahwah, NJ: Erlbaum.

Maskara, A. & Noetzel, A. (1993) Sequence Recognition with Recurrent Neural Networks. *Connection Science, 5*, 139-152.

McLaren, I.P.L. (1993) APECS: a solution to the sequen- tial learning problem. *Proceedings of the Fifteenth Annual Convention of the Cognitive Science Society*, *pp. 717-722.* University of Colorado at Boulder.

McLaren, I.P.L., Green, R.E.A. & Mackintosh, N.J. (1994) Animal Learning and the Explicit/Implicit Distinction. In N.C. Ellis (Ed.), *Implicit and Explicit Learning of Languages.* London: Academic Press.

McLeod, P., Plunkett, K. & Rolls, E.T. (1998) *Introduction to Connectionist Modelling of Cognitive Processes*. Oxford: Oxford University Press.

Nissen, M.J. & Bullemer, P. (1987). Attentional require ments of learning: Evidence from performance measures. *Cognitive Psychology, 19*, 1-32.

Norusis, M.J. (1994) *SPSS Advanced Statistics 6.1.* [Handbook]. Chicago: SPSS Inc.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986). Learning Internal Representations by Error Propagation. In D.E. Rumelhart & J.L. McClelland (Eds.), *Parallel distributed processing*, Vol. 2. Cambridge, MA: MIT-Press.

Spiegel, R. & McLaren, IPL. (2001). Recurrent Neural Networks and Symbol Grounding. *Proceedings of the International Joint INNS/IEEE Conference on Neural Networks*, Washington, D.C.

Spiegel, R., Jones, F.W. & McLaren, IPL. (2001). The Prediction-Irrelevance Problem in Grammar Learning. *Proceedings of the International Joint INNS/IEEE Conference on Neural Networks*, Washington, D.C.

Servan-Schreiber, D., Cleeremans, A. & McClelland, J.L. (1991) Graded State Machines: The representation of temporal contingencies in simple recurrent networks. *Machine Learning, 7*, 161-193.

Timmermans, B. & Cleeremans, A. (2000) Rules versus Statistics in Biconditional Grammar Learning: A Simulation based on Shanks et al. (1997). In L.R. Gleitman & A.K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*. Mahwah, N.J.: Erlbaum.

Effect of Choice Set on Valuation of Risky Prospects

Neil Stewart (neil.stewart@warwick.ac.uk)
Nick Chater (nick.chater@warwick.ac.uk)
Henry P. Stott (hstott@owc.com)
Department of Psychology, University of Warwick
Coventry, CV1 3GA, UK

Abstract

Current models of decision making under risk assume access to the absolute magnitudes of gamble attributes. The two experiments presented here provide evidence that decisions under risk are based, in addition, on the context of the decision. In Experiment 1 the set of options offered as certainty equivalents was shown to determine the value of simple gambles of the form "p chance of £x". Experiment 2 employed a novel procedure where the payment structure was such that it was optimal for participants to provide truthful certainty equivalents. Again, the context provided by the set of certainty equivalents influenced the choice of certainty equivalent.

Many existing theories of decision making under risk (e.g., Kahneman & Tversky, 1979; Quiggin, 1982; Tversky & Kahneman, 1992; von Neumann & Morgenstern, 1947) predict that participants use the absolute magnitudes of value and probability in making risky decisions, or that some monotonic transform of these attributes is used. For example, in expected utility theory (von Neumann & Morgenstern, 1947), the utility of an outcome is a negatively accelerated function of value, and outcomes which maximize the expected utility are preferred. The aim of the experiments described here was to investigate to what extent contextual factors also influence decision under risk.

Mellers, Ordóñez and Birnbaum (1992) measured the attractiveness ratings and buying prices of simple binary gambles presented in two different contexts. In one context, the distribution of expected values of accompanying gambles was positively skewed, and in the other context, the expected values were negatively skewed. Attractiveness ratings were influenced by context. However, for simple gambles of the form "p chance of £x" context had a minimal effect on buying price. With more complicated gambles of the form "p chance of £x otherwise £y", the effect was slightly larger. The effect of context on attractiveness and the lack of an effect of context on buying price is consistent with a similar demonstration by Janiszewski and Lichtenstein (1999), and consistent with a review of previous research by Poulton (1982).

However, context does affect choice of certainty equivalents (hereafter, CEs) in other conditions. (CEs are the amount of money that can be obtained for certain, participants feel is equivalent to a given gamble). Birnbaum (1992) demonstrated that skewing the distribution of CEs offered for simple gambles, whilst holding the range constant, influenced the selection of a CE. When the CE options were positively skewed (i.e., more small values) gambles were over-valued compared to the negatively skewed context, consistent with range-frequency theory (Parducci, 1965; 1974).

The aim of Experiment 1 was to demonstrate that the options offered as potential CEs influence estimates of a gamble's CE. Following a similar logic to a loudness judgment experiment by Garner (1954), participants were given a set of potential CEs for each gamble, and asked to choose the option closest to their estimate of the CE for each gamble. For each gamble, CE options were either all lower in value than the free choice CE (given by another group of participants) or all higher. If participants were not influenced by context, then their choices of CE should be highly skewed towards the mean free choice CE. If participants' responses are solely determined by context, then the distribution of responses across options should be the same for both the low and high value range of CEs. Experiment 2 introduces a new procedure to investigate these context effects in which it is optimal for participants to provide truthful CEs.

Experiment 1

The curve in Figure 1 represents a hypothetical normal distribution of CEs given under free choice conditions. If participants are not affected by the context provided by the range of CEs offered in the restricted choice conditions, it is possible to predict their distribution of responses. For an option at the lower end of the range, the probability of selecting that option is the integral of the free choice function between -• and a point half way between the lowest two options. Similarly for an option at the higher end of the range, the probability is the integral between a point half way between the highest two options and +• . For an option of intermediate value, the probability is the integral between the bound half way between the intermediate option and the next lowest option, and bound half way between that intermediate option and the next highest option. In other words, it is assumed

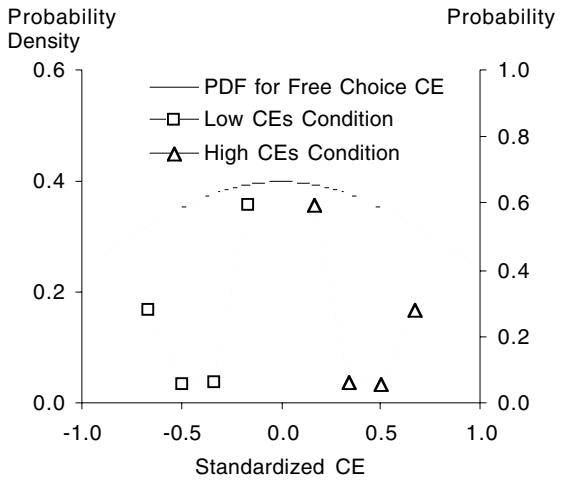


Figure 1: Predicted choices in Stage 2, under the assumption that context will not affect CE choice.

participants choose the option nearest to the CE they would have chosen under free choice conditions.

For the options used in Experiment 1 (see the Design for a description), the two lines in Figure 1 represent the expected distribution of responses. If participants are not influenced by the context provided by the four CE options then the key prediction is that participants in the high CEs condition should choose the lowest option more than half of the time, and participants in the low CEs condition should choose the highest option more than half of the time. This prediction holds for any symmetrical distribution of free choice CEs.

Method

Participants 30 psychology undergraduates from the University of Warwick participated for course credit. 14 took part in Stage 1 of the experiment. 16 took part in Stage 2.

Design Stage 1 was designed to measure participants' free choice CEs for a series of simple gambles. A set of 20 gambles was created by crossing the amounts £200, £400, £600, £800 and £1000 with the probabilities 0.2, 0.4, 0.6 and 0.8. In Stage 2 different participants were presented with the same gambles, and asked to select from a set of four CEs the CE that was closest to their judgment of the value of the gamble. For each gamble two sets of CEs were created. In the low CEs condition participants received options all lower than the mean free choice CE given in Stage 1. In the high CEs condition participants received CEs that were all higher. The CE sets were constructed as follows. The mean and standard deviation of the free choice CEs was calculated for each gamble. The two sets of equally spaced CEs (for the high value and low value conditions) were calculated so that their range was equal to approximately half the free choice standard deviation. One set of CEs was placed below the free choice value, and the other set above. This difference between the lowest CE in the high CEs condition (or the highest CE of the low CEs condition) and the mean free choice CEs was set to be roughly equal to the difference between CEs within a condition. Options were rounded to be familiar, easy to deal with values.

Procedure Participants were given written instructions. They were asked to imagine choosing between "£30 or a 50% chance of £100" to illustrate that gambles could have a value. They were told they would be asked to value a series of gambles, and that they should imagine they had the chance to play the gamble. For each gamble they were asked how much money for certain they thought it was fair for someone to give them for the other person to have a chance to take the gamble instead. They were also asked to consider the opposite situation, where they would be buying the gamble. It was explained that the purpose of the experiment was to investigate how much they thought the gambles were worth, and that there was no correct answer. For participants in Stage 2, it was explained that they should choose the CE option nearest the value they thought the gamble was worth.

Each gamble was presented on a separate page of a 20 page booklet. For participants in Stage 1 gambles were presented as follows:

For you, how much is the gamble
"80% chance of £600"
worth?

£____

Probabilities were always presented as percentages. For participants in Stage 2, a set of options was added. Options were always presented in numerical order, as with the following example of a low CE set:

How much is the gamble
"60% chance of £400"
worth?

Is it: £60 £80 £100 £120

Results

Participants took approximately 5 minutes to complete the task. Figure 2 plots the Stage 1 free choice CE against the gamble amount for the four gamble probabilities. As expected, the average CE increased with both probability of winning and gamble amount demonstrating that participants were sensitive to manipulations of both. The chosen CE was an approximately linear function of the independent effects of gamble amount and gamble probability. Participants were risk averse, with the mean CE being, on average,

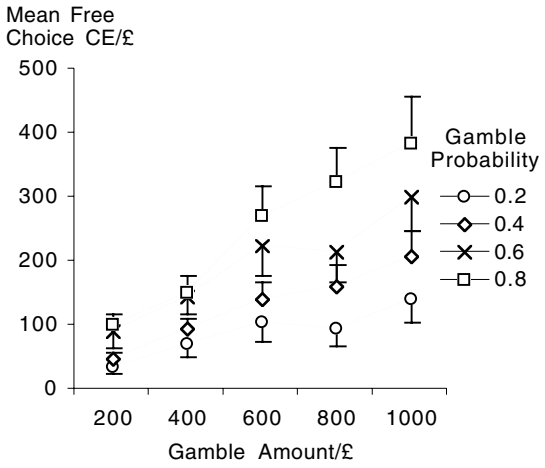


Figure 2: Mean free choice CE as a function of gamble amount for different gamble probabilities for Stage 1 of Experiment 1. (Error bars are standard error of the mean.)

61% of the expected value of the gamble (standard error 3%).

The results of Stage 2 are of most interest. Responses are labeled A through D, with A being the lowest CE, and D being the highest CE. The proportion of times each response type was chosen is plotted in Figure 3. There is no evidence of skewing – instead the distribution of options is approximately the same for the two conditions. The highest CE in the low CEs condition was chosen less than half the time $t(7)=4.21$, $p<0.05$. The same was true of the lowest CE in the high CE condition, $t(7)=5.26$, $p<0.05$.

Discussion

Participants were asked to value a simple gamble of the form "p chance of £x". The effects of p and x on the CE were linear and independent, consistent with Birnbaum's (1992) data. Mean value judgments were below the expected value (£px) of the gamble showing participants were risk averse. Different participants were given a restricted set of CEs for each gamble, either all lower than the mean free choice CE for every gamble or all higher. CE judgments were completely determined by the range of CEs offered, and were not skewed towards the mean free choice CE. Control conditions, not reported here, rule out task demand characteristics as a potential account of these findings, as participants were happy to give highly skewed responses in these conditions. Further, only one of the 16 participants from the second stage reported they were unhappy with the restricted range of CEs offered.

Experiment 2

Experiment 2 was designed to demonstrate the same effect of restricting the range of CEs in a task where it

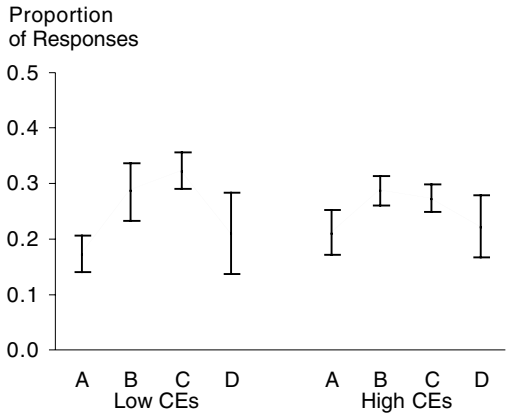


Figure 3: Proportion of responses to each CE for Stage 2 of Experiment 1. (Error bars are standard error of the mean.)

was optimal for participants to report CEs truthfully. This procedure is simpler than other methods used to elicit truthful CEs (e.g., the first price auction, or the Becker, DeGroot & Marschak, 1964, procedure). Specifically, participants divide a given amount into an amount for certain, and an amount to be won with a certain given chance. For example, they might split £1000 into a sure amount of £300 and a "60% chance of £700" gamble. Participants know that the experimenter will select either the gamble or the sure amount, taking the "better" of the two, leaving the participant with the other. Thus it is optimal for participants to split the given amount so that the resulting fixed amount has the same utility as the resulting gamble.

Method

Participants 17 participants took part in the Stage 1 of the experiment. 19 different participants took part in the Stage 2. All participants were paid £4 plus performance related winnings of up to £4.

Design In each trial in Stage 1 a participant divided a given amount of money, £x, into two smaller amounts, £y and £z, to make one fixed amount (£y) and a gamble. There is a given probability p of winning £z, otherwise nothing. Probability p is known to participants before splitting amount £x. Participants know that (if the trial is selected at random at the end of the experiment) the experimenter will take either the fixed amount or the gamble for themselves leaving the participant with the other. It is therefore optimal for the participant to split the amount £x into amounts £y and £z such that £y and a p chance of £z have equal utility for them, i.e., £y is the CE for the gamble "p chance of £z". Under the assumption that the experimenter has the same utility function as them, participants understand that the experimenter will choose the gamble with greater

utility, leaving them with less, if they do not split the amounts in this way.

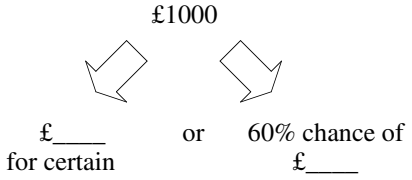
Stage 2 of the experiment differed by offering participants a choice from a set of four pre-split options, rather than giving them a completely free choice. That is, values for £y and £z were presented, and participants selected one pair which could be played at the end of the experiment. As in Stage 1, participants knew that the experimenter would choose the option from the chosen pair with the greatest utility, and therefore they should choose the pair of options closest in expected utility.

It was hypothesized that the pairs of values for y and z presented in Stage 2 would influence participants' choices, and that participants would therefore not just choose the optimal pair. To demonstrate this there was one between participants factor. The set of values for £y and £z were either selected such that £y was always greater than the free choice value of £y from Stage 1 (for equal £x) and £z smaller than the free choice value, or vice versa. The option sets were constructed as follows. The mean and standard deviation of the free choice amount was calculated for each gamble. The two sets of equally spaced options (for the high value and low value conditions) were calculated as described for Experiment 1. As in Experiment 1, if participants are not influenced by the set of choices, then the distribution of responses across the options should be skewed towards the free choice splitting.

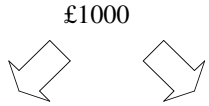
Procedure For both stages the experiment began with instructions. It was explained to participants that they were playing a gambling game, and that they should try to win as much money as possible. They were told the purpose of the experiment was to investigate how much people thought gambles were worth. The task was described. It was emphasized that it was optimal for them to split the money so they thought the amount for certain was equal in worth to a chance on the gamble. They were told that if they allocated funds so either the certain amount was worth more than the gamble, or vice versa, then the experimenter would take the better one, leaving them with less than if they had allocated the money so the gamble was worth the certain amount. They were told that although they could not be certain what the experimenter would do, they should assume the experimenter would behave like them.

Participants were given five practice trials to complete. One of the trials was chosen at random, and it was explained that if the experimenter chose the fixed amount, then the gamble would be played, and they would get the winnings. They were also told that if instead the experimenter took the gamble they would get the fixed amount. Note that this discussion was hypothetical, and participants were not actually told what the experimenter's preference would be.

After the practice the experiment began. The participant completed a booklet of gambles. Gambles were presented in a random order to each participant. An example page from a free choice condition booklet is shown below.



In the restricted choice conditions, pre-split options were presented as in the example below.



			Tick
£322 for certain	or	60% chance of £678	
£334 for certain	or	60% chance of £666	
£346 for certain	or	60% chance of £654	
£358 for certain	or	60% chance of £642	

When the experiment was completed one trial was chosen at random, and played to determine each participant's bonus using an exchange rate.

Results

Participants took between half an hour and one hour to complete the booklet. One participant was eliminated from subsequent analysis for showing a completely different pattern of results to other participants, suggesting they had misunderstood the task. The participant had decreased the value of the fixed amount, £y, as the chance of the gamble amount, p, increased (i.e., they responded as if more likely gambles were worth less to them). 14 out of the remaining 512 trials (16 participants x 32 trials) with incorrect arithmetic were deleted, and treated as missing data.

Figure 4 plots the average fixed amount £y as a function of the gamble chance p for the different total amounts (£x). As the total amount £x increased, then participants' allocation of the fixed amount £y increased. As the probability p of winning the gamble increased participants' estimates of the value of the gamble, £y, also increased. Thus participants' responses seem lawful and sensible. These two effects are approximately independent. The dashed lines in Figure 4 represent risk neutral responding. Data points falling above the dashed line demonstrate risk averse behavior. On average, participants were risk averse for low gamble chances (p=0.2), risk neutral for intermediate gamble chances (p=0.6) and slightly risk prone for high gamble chances (p=0.8). However, standard deviations were approximately 15% of the mean fixed amount allocated and thus for larger gamble chances

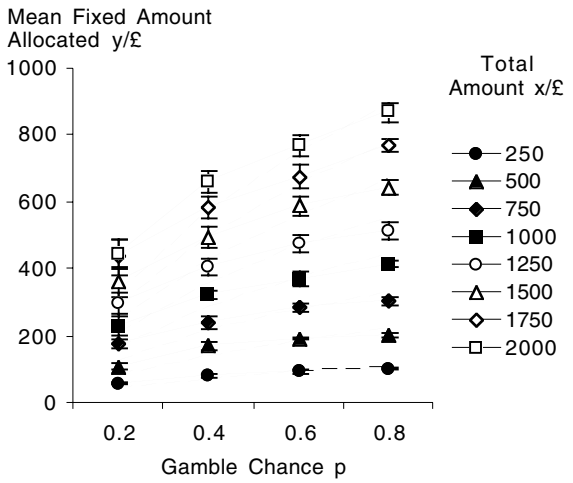


Figure 4: Mean fixed amount allocated in Stage 1 of Experiment 2 as a function of gamble chance for the different total amounts. (Error bars are standard error of the mean.)

approximately half of responses were risk prone, and half risk averse.

The choices made in Stage 2 are shown in Figure 5. Participants did prefer end options over central options in both conditions, consistent with the pattern of results expected if participants were to show no context effect. However, option D in the low y condition was chosen significantly less than half the time, $t(9)=3.47$, $p<0.05$. Similarly, option A in the high y condition was chosen significantly less than half the time, $t(8)=4.20$, $p<0.05$. This observation is consistent with participants showing some context effect. In other words, the proportion of times each option was selected differed significantly from the proportions expected under the assumption that context would not have an effect.

Discussion

The new procedure for eliciting CEs under free choice provides results consistent with Experiment 1. In Experiment 1 participants were, on average, risk averse. Under free choice conditions in Experiment 2 participants were only risk averse for low probability gambles, and were slightly risk prone for high probability gambles. This pattern is the opposite of Tversky and Kahneman's (1986, Problems 10 and 11). Risk averse responding for low probabilities, and risk prone responding for high probabilities is, however, consistent with over estimation of low probabilities and underestimation of high probabilities (e.g., Prelec, 1998; Tversky & Kahneman, 1992; Wu & Gonzalez, 1996). (Consider the case for $p=0.2$. If this probability is overestimated, say at 0.3, then the risk neutral strategy is to increase £y and decrease £z, as the gamble p chance of £z will be overvalued.)

The results of the restricted choice conditions in Experiment 2 replicate those shown in Experiment 1

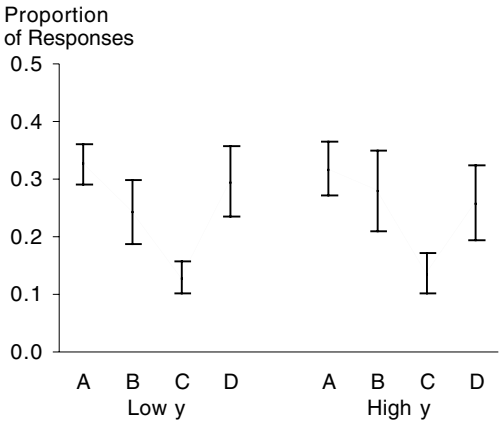


Figure 5: Proportion of each response choice for Stage 2 of Experiment 2. (Error bars are standard error of the mean.)

under a more rigorous procedure, despite participants taking at least six times longer to complete the task compared to Experiment 1. When participants were presented with a range of pre-split total amounts, so that the CE options were either always lower or always higher than the free choice value, their choice of CE was not skewed towards the mean free choice CE. The context provided by the pre-split options influenced their choice of CE.

General Discussion

In the experiments presented here, participants' choice of CE for simple gambles was affected by the range of option CEs offered to them, compared to CEs given by different participants under free choice conditions. For example, when the option CEs were all lower than the free choice CE, participants behaved as if their CE was lower.

Judged and Choice Certainty Equivalents

Careful discussion by Luce (2000) highlights the a difference between judged CEs, where participants provide a single judgement of the value of a gamble, and choice CEs, derived from a series of choices between gambles and fixed amounts. For example, for the kinds of gambles used here, with large amounts and moderate probabilities, judged CEs are overvalued compared to choice CEs (e.g., Bostic, Herrnstein & Luce, 1990). Luce (2000) advocates developing separate theories to for judged and choice certainty equivalents. Participants in these experiments were instructed to complete the restricted CE conditions by judging CEs. However, in Experiment 2, the design should certainly have encouraged imagining choices between gambles and fixed amounts. The degree to which this was the case may explain the 'u' shaped

pattern of results in Experiment 2, rather than the 'n' shaped pattern in Experiment 1.

Other Context Effects

Here the context provided by a set of certainty equivalent options has been found to influence the CE. In other experiments, the context provided by a set of gambles has been shown to influence preferences amongst those gambles. For example, Simonson and Tversky (1992) demonstrated that there was a tendency to prefer a given option when there are other options in the choice set that are unfavorable when compared to the given option. Specifically, in making risky choices between three three-outcome gambles, a gamble was preferred if it dominated another gamble in the choice set.

Implications

Existing models of decision making under risk typically assume that only the attributes of the gamble need be considered when reaching a CE decision. The context or anchoring effects demonstrated here show that the context also needs to be considered. The extent to which context can cause deviation from 'rational choice' has implications for other domains, such as economics and political science, where 'rational choice' models of the individual are applied (e.g., expected utility theory and game theory).

Relation to Perception

The demonstration of context effects in risky decision making suggests that the representation of the utility dimension is similar to that for perceptual psychological dimensions where context effects have also been demonstrated. For example, Garner (1954) showed that participants were completely unable to determine which of six tones was more or less than half as loud as a reference loudness. Instead, participants' judgments were entirely influenced by the range of the six tones. (Laming, 1997, provides an extensive discussion of other similar findings.) Further research is underway in this laboratory to investigate to what extent the context provided by simultaneously presented gambles (see also Mellers et al., 1992) and the context provided by recently considered gambles affects the utility of gambles. This research should help to establish whether utility really is like perceptual dimensions.

Acknowledgments

This work was supported by a grant from Oliver, Wyman & Company to the Institute of Applied Cognitive Science, Department of Psychology, University of Warwick. We are grateful to Gordon D. A. Brown for his insightful comments.

References

Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. Behavioral Science, 9, 226-232.

Birnbaum, M. H. (1992). Violations of monotonicity and contextual effects in choice-based certainty equivalents. Psychological Science, 3, 310-314.

Bostic, R., Herrnstein, R. J., & Luce, R. D. (1990). The effect of preference-reversal phenomenon of using choice indifferences. Journal of Economic Behavior and Organization, 13, 193-212.

Garner, W. R. (1954). Context effects and the validity of loudness scales. Journal of Experimental Psychology, 48, 218-224.

Janiszewski, C., & Lichtenstein, D. R. (1999). A range theory account of price perception. Journal of Consumer Research, 25, 353-368.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47, 263-291.

Laming, D. (1997). The measurement of sensation. London: Oxford University Press.

Luce, R. D. (2000). Utility of gains and losses: Measurement-theoretical and experimental approaches. Mahwah, NJ: Erlbaum.

Mellers, B. A., Ordóñez, L. D., & Birnbaum, M. H. (1992). A change-of-process theory for contextual effects and preference reversals in risky decision making. Organizational Behavior and Human Decision Processes, 52, 311-369.

Parducci, A. (1965). Category judgment: A range-frequency theory. Psychological Review, 72, 407-418.

Parducci, A. (1974). Contextual effects: A range-frequency analysis. In L. Carterette and M. P. Friedman (Eds.), Handbook of Perception (Vol. II). New York: Academic Press.

Poulton, E. C. (1989). Bias in quantifying judgments. Hillsdale, NJ: Erlbaum.

Prelec, D. (1998). The probability weighting function. Econometrica, 66, 497-527.

Quiggin, J. (1993). Generalized expected utility theory: The rank-dependent model. Boston: Kluwer Academic.

Simonson, I., & Tversky, A. (1992). Choice in context: Tradeoff contrast and extremeness aversion. Journal of Marketing Research, 29, 281-295.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty, 5, 204-217.

von Neumann, M., & Morgenstern, O. (1947). Theory of games and economic behavior. (2nd ed.). Princeton, NJ: Princeton University Press.

Wu, G., & Gonzelez, R. (1996). Curvature of the probability weighting function. Management Science, 42, 1676-1690.

The Fate of Irrelevant Information in Analogical Mapping

C. Hunt Stilwell (stilwell@psy.utexas.edu)
Department of Psychology, University of Texas, Mezes Hall 330
Austin, TX 78712

Arthur B. Markman (markman@psy.utexas.edu)
Department of Psychology, University of Texas, Mezes Hall 330
Austin, TX 78712

Abstract

Research on analogical mapping has not yet focused on the fate of information about the base and target domains that is not relevant to the correspondence. We suggest that there are two methods for dealing with irrelevant information in analogies. Nonalignable objects are ignored, while irrelevant attributes of alignable differences are packed away using a process that is a long term equivalent to the suppression process observed in the text comprehension literature. We report one study that supports this hypothesis by demonstrating that unpacking irrelevant information interferes with memory for domains involved in a comparison.

Introduction

Analogical reasoning allows people to compare across domains that might not seem similar on the surface (Gentner, 1983; Holyoak & Thagard, 1995; Keane, Ledgeway, & Duff, 1994). For example, the atom is like the solar system, because something (electrons and planets) revolves around something else (the nucleus and the sun) in each. An open question in analogy research is what happens to the knowledge that is not relevant to this similarity (e.g., electrons are very small and planets are very large)? Most research on comparisons and analogies has focused only on the relevant information. However, the fate of irrelevant information such as the size of the orbiting object in the above example can have important implications for models of analogy and comparisons in general.

In order to place this issue in context, we first discuss the structural alignment process. We then propose a mechanism for dealing with irrelevant information, called *packing*, and compare it to a similar mechanism in language comprehension. Finally, we report an experiment testing this mechanism.

Irrelevant information in analogy

The structural alignment process is used to compare two domains in an analogy. The process operates over structured representations in which the relations between objects are explicit. This leads to two types of information that can be used in the analogy: relational information and object information. Relational information is simply information about the relations between objects, while object information includes the relations the object participates in and the attributes of the object. In an analogy, the two domains are aligned so that their common relational structures are placed in correspondence, and this leads to relational information becoming focal (Markman & Gentner, 1997).

In structural alignment's original formulation, only relational information was used in the comparison of domains (Gentner, 1983, 1989). However, when the theory was extended to ordinary similarity comparisons (Gentner & Markman, 1997; Markman & Gentner, 1993; Medin, Goldstone, & Gentner, 1993), object attribute information became relevant. In a similarity comparison, if the object match is better than the relational match, then it will be preferred. There is also evidence that attribute information influences analogical comparisons. For example, analogs are easier to retrieve if they share attribute similarity to the target than when they share only relational similarity (Gentner, Rattermann, & Forbus, 1993). The fact that attribute information is available in analogical comparisons raises an important question: How does structural alignment deal with attribute information in analogies, where it is not relevant?

One way in that superfluous attribute information might adversely affect the comparison process is by taxing working memory. Recently the role of working memory in analogy has become a topic of study (Hummel & Holyoak, 1997; Waltz, et al., 2000). One finding from this work is that straining working memory hinders the discovery of common relational structures. This is because relational matches take up more working memory capacity than do surface (attribute) matches. Thus anything that decreases the working memory load (e.g., making attribute information less available) will facilitate relational comparisons.

Structural alignment handles some irrelevant information through the concept of alignability. Alignable objects, or objects that participate in the relational correspondence, are relevant to the comparison. In analogies, these objects are generally *alignable differences*, i.e., nonidentical objects that are

placed in correspondence by virtue of playing a common role in a matching relational structure. Alignable differences can be contrasted to *nonalignable differences*, which are objects that do not participate in the common relational structure. Alignable differences are more focal to comparisons than are nonalignable differences.

In this paper we are interested in a second type of irrelevant information that has not received much attention in the analogy literature. In particular, when an alignable difference is found, some of the attributes of the corresponding objects are likely to be irrelevant to the relational match. For example, in Figure 1, the pig in the top scene and the baby in the bottom scene are an alignable difference, because each is making a mess. Attribute information about the pig such as its snout and ears are not relevant to the relational match. Is information like this treated as focal to the comparison because it is part of an alignable difference, or is it treated like a nonalignable difference because it is irrelevant to the relational correspondence?

Packing and Suppression

The comparison process might deal with irrelevant attribute information by packing it away, or making it less available for processing. On this view, when a representation is packed during a comparison, the representations involved are changed so that only the information relevant to the match between domains is immediately available.

A similar mechanism—called *suppression*—has been suggested in the language comprehension literature. Suppression inhibits superfluous information about a word or concept during comprehension. For example, in the sentence "He won the match," borrowed from Gernsbacher & Robertson (1999), the inappropriate meaning of "match as a stick used to light a fire is inhibited. This mechanism is also useful for limiting the amount of information in working memory during processing. Nonetheless, there are key differences between the processes of analogical comparison and language comprehension that might limit the utility of a suppression mechanism in analogical reasoning. For one, language comprehension - and therefore suppression - occurs very rapidly, while analogies form over a longer period of time. In addition, suppression is short-lived, because the suppressed meaning of a word might be needed (and therefore accessed) in subsequent sentences. In contrast, packing involves representational change, and so we would expect longer-term effects than those observed with suppression.

If irrelevant information about a comparison is packed away, then attributes of alignable differences that are irrelevant to the relational match should be treated like nonalignable differences in comparisons. We tested this possibility by extending a previous study

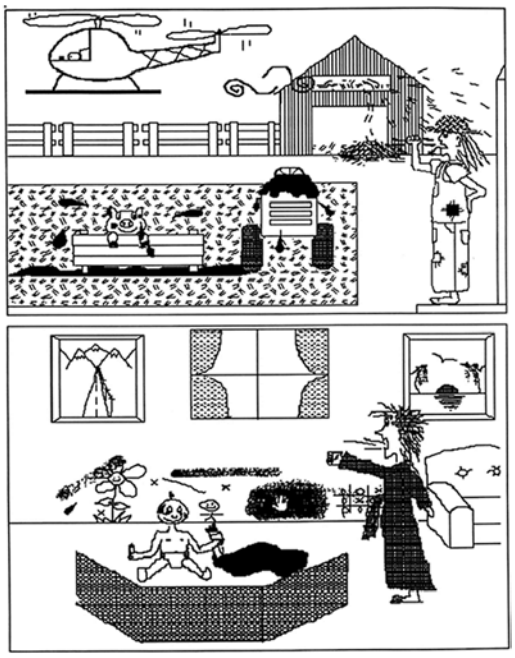


Figure 1: Example Scene Pair. In the top scene the pig is an alignable difference and the helicopter is a nonalignable difference.

by Markman & Gentner (1997). Their study showed that nonalignable differences make poor recall cues, while alignable differences make effective recall cues. In this experiment, subjects were shown pairs of scenes like the one in Fig. 1, and were asked to rate their similarity. As in Fig. 1, there was a *target scene* on top, which contained both alignable differences (e.g., the pig) and nonalignable differences (e.g., the helicopter) with the *comparison scene* on the bottom. After a filler task, subjects were given pictures of objects that were either alignable differences (the pig) or nonalignable differences (the helicopter), and they were asked to recall as much as they could about the scene in which the object had originally appeared. Subjects recalled more information when given the alignable cues than when given the nonalignable cues. This finding suggests that alignable differences are more focal than nonalignable differences.

What aspect of the alignable object is making it a good retrieval cue? The packing mechanism we propose suggests that the connection of the object to relational information is important for retrieval, and that the attributes of the object that are irrelevant to the relational match are packed away. If these attributes are made more salient, the alignable object may be treated more like the nonalignable object, and its efficacy as a recall cue will decrease. To test this idea we added an *unpacking task* in between the comparison and recall tasks of the Markman and Gentner (1997) experiment. In the unpacking task subjects were shown either the alignable or nonalignable difference from the

original scene, and were asked to describe what the object looked like. Listing properties of the objects should cause subjects to focus on the attribute information. If the comparison process made this information less available by packing it away, then when this object is later used as a retrieval cue, it should be ineffective. In particular, the encoding specificity principle suggests that the likelihood of retrieving an item in memory increases with the similarity between the context at retrieval and the context at encoding (Tulving & Thompson, 1973). Thus, if the unpacking task focuses subjects' attention on irrelevant information that was packed away during comparison, the alignable object should no longer be an effective recall cue. In addition, because nonalignable objects are not emphasized by the comparison process, unpacking them should not have any effect on their efficacy as recall cues.

In addition to the unpacking task, we added a packing task in between the comparison and recall tasks. This task was designed to reinstate the analogical mapping. First, during the initial comparison, people were asked to label each pair. This label generally referred to the relational match between the pictures. Later, subjects were asked to recall the titles. Recalling the titles should reinstate the relational mapping for the pair of scenes given that title, resulting in a pattern of recall similar to that observed by Markman and Gentner (1997), with alignable objects serving as effective cues and nonalignable differences as poor recall cues.

In pilot experiments conducted with the packing and unpacking tasks, the results fit with these predictions. When given the unpacking task, alignable cues were no longer effective recall cues. In contrast, when given the packing task, alignable objects retained their efficacy as recall cues. In the present experiment, each subject completed both the packing and unpacking tasks before recall. We predict that the tasks that subjects do last (before recall) should have the most influence on their ability to recall the original scenes. If subjects complete the unpacking task last, the unpacked alignable cues should be no better than the nonalignable cues because attention is focused on the irrelevant attributes. In contrast, if they complete the packing task last, and thus reinstate the relational mapping between scenes, the alignable cue should be a better recall cue than the nonalignable cue.

Method

Participants

Subjects were 172 undergraduates at the University of Texas, Austin, who participated for course credit. The data from 44 subjects was not used due to their failure to follow directions. Most of these either listed properties during the recall task or failed to complete one of the tasks. This left 128 subjects for analysis.

Procedure

The procedure is summarized in Figure 2. As this is a between subject design, each subject saw the tasks in the order specified by either the right or left column in this figure.

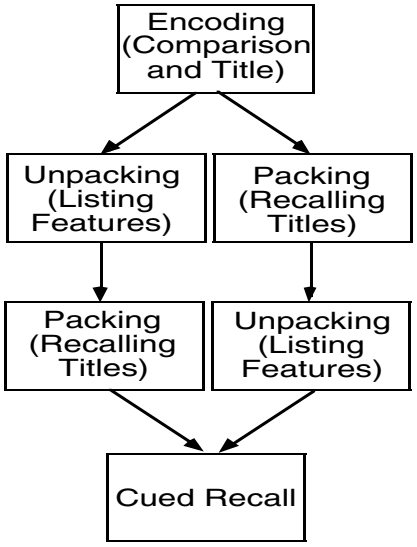


Figure 2: Design of the Experiment.

Subjects sat in cubicles and performed the experiment at their own pace. They were instructed to fill out a set of packets in order, from top to bottom, without looking back or ahead. The Comparison Task packet, which was always completed first, instructed subjects to look at each pair of scenes and rate their similarity on a nine-point scale. They were also instructed to write a descriptive title for the pair of scenes at the bottom of the page after rating their similarity. Participants took approximately 5-10 minutes to complete this task. Subjects then completed an unrelated filler task that took approximately 15-20 minutes.

After the filler task, subjects were given either the Packing Task packet or the Unpacking Task packet. In the Packing Task, subjects wrote down as many of the titles that they had given the scene pairs as they could remember. In the Unpacking Task, subjects wrote down as many properties of the objects, which had been either alignable or nonalignable objects in the comparison scenes, as possible on the lines provided. Each subject completed both the Packing and Unpacking Task. They then completed another unrelated filler task, which took approximately 15-20 minutes.

Finally, subjects were given the cued-Recall Task packet. They were told that they would see a series of objects that had appeared in the scenes they had seen earlier. They were instructed to write down as much as they could remember about the scenes in which the objects had originally appeared. These objects were

either the alignable or nonalignable objects from the comparisons scenes, as in the Unpacking Task.

Design

The study used a 2(Task Order: Packing Last vs. Unpacking Last) X 2(Unpacked Object: Alignable vs. Nonalignable) X 2(Recall Object: Alignable vs. Nonalignable) design. All ten items were run in all conditions. Task order was between subjects. A total of 8 subjects was required to get one observation for each item in each condition.

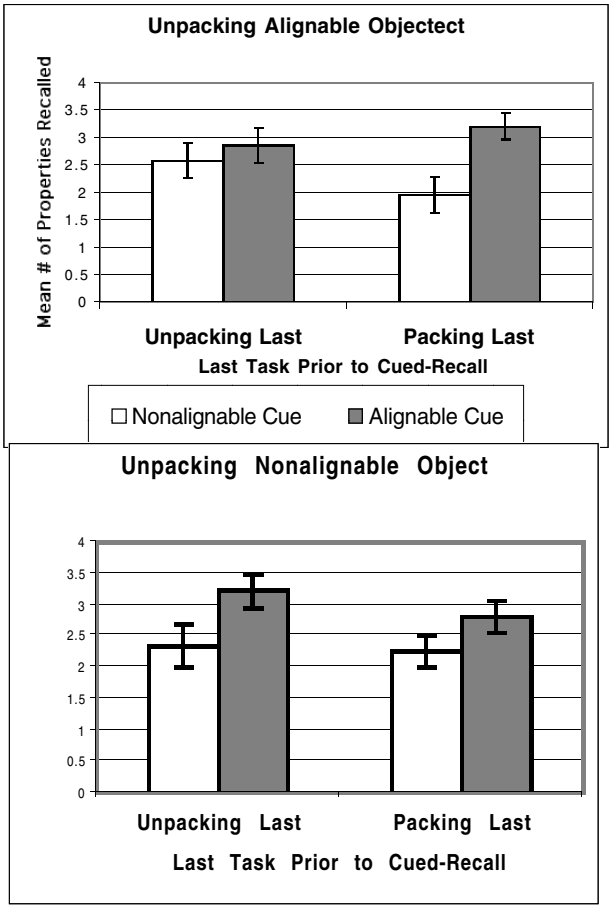


Figure 3a & 3b: Mean number of Properties Recalled.

Results

Figure 3a shows the pattern of recall for subjects who unpacked the alignable object during the Unpacking task. Figure 3b shows the recall data for subjects who unpacked the nonalignable object.. A repeated measures 2(Task Order) X 2(Unpacking) X 2(Recall) ANOVA was performed, by item, on the mean number of properties recalled. There was a main effect of recall, $F(19,1) = 6.33$, $p < .05$. This reflects the fact that alignable cues are generally better than nonalignable cues. There was also a significant Task Order by Recall interaction, $F(19,1) = 2.41$, $p < .05$. As predicted, there was a significant Task Order by Unpacking by Recall interaction, $F(19, 1) = 8.76$, $p < .05$.

The critical predictions for this study center on the conditions in which the alignable difference was unpacked. When the unpacking task was done last (i.e., most recently before recall), we expected this task to interfere with recall. In this case, the alignable difference should not be a good retrieval cue. In contrast, when the packing task was done last, subjects should be able to reinstate the analogical mapping. In this case, the alignable difference should be a much better retrieval cue than the nonalignable difference. Consistent with this prediction, when subjects completed the unpacking task last and unpacked alignable objects, the difference between alignable vs. nonalignable recall cues was not significant, $t(38) = .27$, $p > .10$. However, as predicted, when the packing task was completed last, the alignable cue was a better recall cue than the nonalignable cue, for subjects who unpacked the alignable object, $t(38) = 3.03$, $p < .05$. These data are shown in Figure 3a.

A different pattern of data was obtained when the nonalignable difference was unpacked. In this case, unpacking should not affect the efficacy of the alignable difference as a retrieval cue. Consistent with this prediction, subjects recalled more differences given the alignable difference cue than given the nonalignable difference cue regardless of whether the nonalignable difference was unpacked before or after the packing task was performed. These data are shown in Figure 3b.

Further tests were performed taking into account whether subjects remembered the title for the scene. This analysis is important for determining if it was the reinstatement of the relational mapping through recalling the title that resulted in the different patters of recall. For this analysis an additional factor, Title Recall, was added to the previous analysis. In the packing last condition, subjects who unpacked the alignable object and received the alignable recall cue recalled significantly more properties of the original scene when they recalled the title ($m = 3.97$) than when they did not ($m = 2.48$), $t(38) = 2.36$, $p < .05$. The same was true for subjects in the unpacking last condition. They recalled more when they remembered the title ($m = 3.47$) than when they did not ($m = 2.08$), $t(38) = 2.63$, $p < .05$. This finding suggests that successfully recalling the title of a pair, and thus reinstating the mapping, was beneficial for later recall.

We also looked at whether subjects recalled at least one thing about the original scene in the recall task. This tells us whether they were able to retrieve the original representation of the scene. The mean

proportions of subjects recalling at least one property of the scene are presented in Table 1. The proportions data shows the same pattern that the mean recall data showed. A 2(Task Order) X 2(Unpacking) by 2(Recall) ANOVA was performed on the proportions data, and showed the same results as the mean recall data. There was a main effect of recall, $F(19, 1) = 18.92$, $p < .05$, again showing that alignable cues are generally better than nonalignable cues at facilitating recall. Again, there was also an Order by Recall interaction, $F(19, 1) = 6.35$, $p < .05$. Most importantly, the Order by Unpacking by Recall interaction was significant, $F(19, 1) = 9.22$, $p < .05$. Thus the results exhibited the same pattern both for total amount of recall as well as proportion of subjects who recalled anything about a scene. This finding suggests that the packing task facilitates access to the original mapping rather than increasing the availability of additional properties of the scene.

Table 1: : Proportion of trials on which one or more properties were recalled

	Unpacking Alignable Object	
	Alignable Cue	Nonalignable Cue
Packing Last	0.70	0.39
Unpacking Last	0.58	0.51
	Unpacking Nonalignable Object	
	Alignable Cue	Nonalignable Cue
Packing Last	0.63	0.47
Unpacking Last	0.66	0.46

Discussion

This study provides evidence that information about alignable differences that is not relevant to the relational match is packed away during comparison. As expected, when people unpacked information about the alignable difference by listing properties of it, the efficacy of the alignable difference as a retrieval cue was reduced. Nonalignable differences, which are not focal in comparison, were not influenced significantly by the unpacking task. These data did not simply reflect a general interaction between the packing task and the retrieval task, because performing the packing task (in which subjects recalled titles they had given to the pair) restored the efficacy of the alignable difference as a recall cue.

It is straightforward to view the packing task as increasing the salience of properties of the alignable differences that had been packed away. It is less clear what is happening during the packing task. We suggest that recalling the title causes subjects to reinstate the relational mapping. One piece of evidence in favor of this interpretation is that recalling the title provides a significant boost in the level of recall. One aspect of the data that bears further scrutiny is the observation that successfully recalling the title of a pair increased the efficacy of the alignable cue regardless of whether the unpacking task was done before or after unpacking the alignable object.

This experiment raises several questions about the packing phenomenon that can be addressed in future research. First, how long lasting are the representational changes that occur when a representation is packed or unpacked? The results of this experiment demonstrate that the effects of packing or unpacking a representation last at least 10-15 minutes. This time course contrasts with the suppression mechanism (Gernsbacher & Robertson, 1999), which lasts a much shorter period of time. Future research could examine the effects of these tasks over longer delays.

A related question involves how packing or unpacking a representation affects future comparisons using that representation. One possibility is that the role of surface similarities, which have been shown to figure prominently in the early stages of comparisons (Goldstone & Medin, 1994; Ratcliff & McKoon, 1989) might be attenuated if packing makes these surface features less available than relational information.

This research is an initial step toward exploring the fate of irrelevant information in comparisons. The packing phenomenon is likely to be useful in guiding future research. It is consistent with the current interest in the role of working memory in analogical mapping, and may prove useful in answering questions about how relational information and working memory interact. In addition, it may provide valuable constraints for current models of analogy, which typically focus selectively on the relevant information in a comparison.

Acknowledgments

This research was supported by NSF grant SBR-9905013 given to the second author. The authors thank Eric Dietrich for discussions that laid the theoretical foundation for this work. They thank Sam Day, Dedre Gentner, Ryan Gossen, and Gregory Murphy for contributing to the ideas in this paper. Finally, the authors thank Kerry Collins for his performance all season (except in that last game).

References

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2), 155-170.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, 52(1), 45-56.

Gernsbacher, M. A., & Robertson, R. R. (1999). The

role of suppression in figurative language comprehension. *Journal of Pragmatics*, 31(12), 1619-1630.

Goldstone, R. L., & Medin, D. L. (1994). Time course of comparison. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 29(1), 29-50.

Holyoak, K. J., & Thagard, P. (1995*). Mental leaps: Analogy in creative thought.* Cambridge, MA:MIT Press.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104(3), 427-466.

Keane, M. T., Ledgeway, T., & Duff, S. (1994). Constraints on analogical mapping: A comparison of three models. *Cognitive Science*, 18(3), 387-438.

Markman, A. B., & Gentner, D. (1997). The effects of alignability on memory. *Psychological Science*, 8(5), 363-367.

Medin, D.L., Goldstone, R.L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*(2), 254-278.

Ratcliff, R., & McKoon, G. Similarity information versus relational information: Differences in the time course of retrieval. *Cognitive Psychology*, 21(2), 139-155.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80, 352-372.

Waltz, J. A., Lau, A., Grewal, S. K., & Holyoak, K. J. (2000). The role of working memory in analogical mapping. *Memory and Cognition* , 28(7), 1205-1212.

Visual Expertise is a General Skill

Maki Sugimoto (mxs@hnc.com)
HNC Software, Inc.
5935 Cornerstone Court West
San Diego, CA 92121-3728 USA

Garrison W. Cottrell (gary@cs.ucsd.edu)
UCSD Computer Science and Engineering
9500 Gilman Dr.
La Jolla, CA 92093-0114 USA

Abstract

The fusiform face area (FFA) in the ventral temporal lobe has been shown through fMRI studies to selectively respond with high activation to face stimuli, and has been identified as a face specific processing area. Studies of brain-lesioned subjects with face recognition or object recognition deficits also have often been cited as evidence for face specific processing. Recent studies, however, have shown evidence that the FFA also responds with high activation to a wide variety of non-face objects if the level of discrimination and the level of expertise are controlled. Based on these recent results, we hypothesized that the features of faces that the FFA respond to can be useful for discriminating other classes of visually homogeneous stimuli with some tuning through experience. To test our hypothesis, we trained two groups of feed-forward neural networks on visual classification tasks. The first group was pretrained on basic level classification of four stimulus classes, including faces. The second group was pretrained on subordinate level classification on one of the stimulus classes and basic level classification on the other three. In two experiments that used different criteria to stop pretraining, we show that networks that fully acquire the skill of subordinate level classification consistently show an advantage in learning the new task.

Introduction

The functional role of the so-called fusiform face area (FFA) located in the ventral temporal lobe is controversial. The FFA has been shown in fMRI studies to respond with high activation to face stimuli but not to other visual object stimuli, and has thus been identified as a face specific processing area (Kanwisher, 2000). Studies of patients with face recognition or object recognition deficits also have often been cited as evidence for face specific processing. Recent studies by Gauthier and colleagues have questioned whether the FFA is really a face-specific area (Gauthier, Behrmann & Tarr, 1999a). They have proposed an alternative theory that the FFA engages in expert level classification of visually similar stimuli from a wide variety of categories not limited to faces.

The current study is an attempt to shed light on the debate through simulations of computational models. We constructed our hypothesis based on the recent view that the FFA is a domain-general processing area, specializing in visual expertise of fine level discrimination of homogeneous stimuli. Our experimental results show strong support for the hypothesis, thus providing further evidence for the plausibility of the domain-general view of the FFA.

In the following sections, we will describe the evidence for and against the FFA's face specificity, and our refinement of the domain-general hypothesis. The experimental methods are then described in detail followed by the results. We conclude by summarizing our findings and suggesting future research.

Evidence for the Face Specificity of the FFA

Studies of brain-lesioned subjects provide the strongest evidence for localized face specific processing. Patients with *associative prosopagnosia* reportedly have deficits in individual face identification, but are normal in face detection and object recognition (Farah, Levinson & Klein, 1995). On the other hand, patients with *visual object agnosia* are normal in identifying individual faces but have deficits in recognizing non-face objects (Moscovitch, Winocur & Behrmann, 1997). The two groups of patients serve as evidence for a double dissociation of visual processing of faces and other objects.

Through fMRI studies of normal brains, the FFA has been identified as the area being most selective to faces (Kanwisher, McDermott & Chun, 1997). Prosopagnosia patients usually have a lesion in an area encompassing the FFA (De Renzi et al., 1994), providing consistent evidence for the face specificity of the FFA.

Evidence against the Face Specificity of the FFA

Gauthier and colleagues argued that the FFA showed high activity in response to various classes of visual stimuli when the levels of discrimination and expertise were properly controlled (Gauthier et al., 1999a). One study showed significantly high activity of the FFA for car and bird experts when stimuli from their respective expert class were presented (Gauthier et al., 2000). Another study that utilized 3-D artificially rendered models called "Greebles" (Gauthier & Tarr, 1997), showed the FFA increasing its activation in response to the rendered models as the subjects were trained to classify them at a fine level (Gauthier et al., 1999b). For the latter study, the use of the Greebles allowed the authors to develop human subject experts of non-face objects while fully controlling the subjects' experience with the stimuli.

These results showing high activity of the FFA for non-face objects including completely novel objects,

serve as strong evidence against the face specific view of the FFA.

Our Approach with Computational Models

Why does the FFA engage in expert classification of non-face objects? We hypothesized that the features of faces that the FFA responds to can be useful for discriminating any class of visually homogeneous stimuli with some tuning through experience. If our hypothesis is correct, possession of expertise with faces should facilitate the expert level learning of other classes. In this paper, we consider individuating members of a homogeneous class (subordinate classification) to be an expert level task.

To test our hypothesis, we trained two groups of neural networks with hidden layers to perform a subordinate level Greeble classification task. Prior to training on the Greebles, we pretrained the networks on one of the following two tasks:

1. Basic level classification of faces and objects
2. Subordinate level classification of one of the classes and basic level classification of the rest

Developing the first visual expertise for non-face objects is one of the conditions that cannot be ethically achieved in human experiments. Our computational model attempts to overcome this limitation by pretraining neural networks on subordinate classification of non-face objects. If the advantage can be observed for all groups of networks with various pretraining tasks, we would conclude that the features that are discriminative of homogeneous visual stimuli *in general* are robust features that translate well to any other class of stimuli.

Experimental Methods

As described briefly in the previous section, we trained neural networks on subordinate level classification with various pretraining tasks. In this section, we will describe further details on the input database, the preprocessing procedure, network configurations and the simulation procedures.

Image Database

The images were 64x64 8-bit grayscale images consisting of five basic classes: human faces, books, cans, cups, and Greebles. Each class included 5 different images of 12 individuals, resulting in a total of 60 images for each class. Example images are shown in Figure 1 and 2. The non-Greeble images are described elsewhere (Dailey & Cottrell, 1999). For the Greebles, the 12 individuals were selected exclusively from one of the five families. Five images were obtained for each individual by performing random operations of shifting up to 1 pixel vertically and horizontally, and rotating up to 3 degrees clockwise or counterclockwise in the image plane. A region from the background of the common object images was randomly extracted and applied to the background of the Greeble images.

Figure 1: Example of face and common object images (Dailey & Cottrell, 1999)

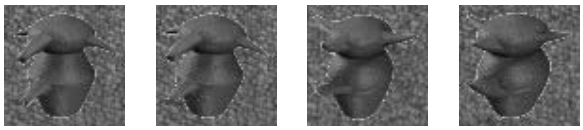


Figure 2: Example of Greeble images;The left two are images of the same Greeble

Preprocessing

To preprocess the images, we followed the procedures introduced by Dailey and Cottrell (1999), applying Gabor based wavelet filters and using principal component analysis (PCA) for dimensionality reduction.

2-D Gabor wavelet filters, which are relatively robust to variations in background, translation, distortion and size (Lades et al., 1993), have been used previously for face recognition tasks with neural networks. Each image was represented by the magnitude of the responses of 40 filters tuned to 8 orientations and 5 spatial frequencies, measured at 64 points subsampled in an 8x8 grid, resulting in a vector of 2560 elements (Buhman, Lange & von der Malsburg, 1990; Dailey & Cottrell, 1999).

PCA was done separately on each spatial frequency, extracting 8 components for each of the 5 scales to form 40-dimensional input vectors. Each element of the input vectors were normalized across all face/object images by z-scoring, i.e., a linear transformation to mean 0 and standard deviation 1. The Greeble patterns were not represented in the principal components to prevent any knowledge of the Greebles contaminating the model.

Network Configuration

Standard feed forward neural networks with a 40-unit hidden layer were used for all the experiments. The hidden layer units used the logistic sigmoid function while the output units were linear. The learning rate and momentum were .005 and .5, respectively. These parameters were tuned so that the networks reliably learned the most difficult task, which was the subordinate level classification on faces and Greebles with basic level classification on the common objects.

Training Set Variations

Each network was trained on a subset of the whole data set as follows. For the classes on which subordinate level

classification were performed, one image for each individual was randomly selected to test generalization. Another image was removed to be used as the holdout set (for early stopping) from the rest of the images, resulting in a reduced training set of 3 images per individual.

For the classes on which basic level classification were performed, images of one randomly selected individual were reserved for testing. Images of a different individual were used as the holdout set, resulting in a reduced training set of images of 10 individuals.

With the arrangements mentioned above, 3 images of 12 individuals were available for use as the training set for the subordinate level classification task and 5 images of 10 individuals were available for the basic level task. In order to control the number of images presented to the networks during the training, the training set was restricted to 3 images from 10 individuals for both levels of classification, for a total of 30 images for each class.

In the experiments reported below, we do not use the holdout and test sets, as we use RMSE thresholds and amount of training as conditions for stopping training phases. The holdout and test sets are used in preliminary experiments to find appropriate values of the RMSE thresholds.

Task Variations

Training of the neural networks was done in two phases. In the first phase, the pretraining phase, the networks were trained using only the face/common object data on one of the following two tasks:

1. Basic level classification on all 4 input categories
2. Subordinate level classification on 1 category and basic level on the rest.

The networks that were assigned the first task had 4 outputs, corresponding to book, can, cup, and face. We will refer to these networks as "Non-experts".

The networks that were assigned the second task had 13 outputs; 3 for the basic level categories and 10 for the individuals in the subordinate level. For example, if a network was assigned a subordinate level classification task for cans and basic level for the rest, the output units corresponded to book, cup, face, can 1, can 2, can 3, etc. We will refer to these networks as "Experts".

In the second phase, the pretrained networks were trained on a subordinate level classification task of individuating Greebles in addition to the pretrained task. Greebles were included in the input data set and 10 output units corresponding to each Greeble were added. Thus, the networks performed either a 14-way or a 23-way classification depending on their pretrained task.

We ran two sets of experiments using different criteria to determine when to stop pretraining:

Experiment 1 The networks were trained until the training set RMSE dropped below a fixed threshold.

Experiment 2 The networks were trained for a fixed number of epochs.

The first criterion controls the networks' familiarity with the input data with respect to their given tasks. This criterion is partly motivated by Gauthier et al.'s definition of experts that takes into account not only the classification accuracy but also the response time which reflects the subjects' degree of certainty. Response time is often modeled in neural networks by the RMSE on a pattern. The second criterion controls the number of opportunities the networks can learn from the input. Employing this criterion corresponds to the idea of controlling the subjects' experience with their tasks, which is often difficult to control in human subject experiments.

For the Greeble phase, the networks were trained to a fixed RMSE threshold for both experiments.

Provided that the networks adequately learned the pretraining task in the pretraining phase, any difference in the learning process of the new task (in the second phase) between the Non-experts and the Experts must be due to the differences in the pretraining task. For the first experiment, we set the pretraining RMSE threshold to 0.0806. This threshold value was determined through preliminary experiments by estimating the training set RMSE for the face expert task to be learned without overfitting. For the second experiment, the epoch limits ranged over $5 * 2^n$ with $n \in 0, 1, .., 10$ to fully analyze the effect of the pretraining task differences. We set the RMSE threshold for the second (Greeble) phase to 0.158. This was determined from similar preliminary experiments based on the estimated optimal RMSE on the most difficult task, subordinate level classification on faces and Greebles.

Evaluation

For the two experiments, we compared the number of epochs required to learn the new task for the Non-experts and the Experts. For experiment 1, we trained 20 networks with different initial random weights for all 5 pretraining tasks, for a total of 100 networks. For experiment 2, we trained 10 networks with different initial random weights for all 5 pretraining tasks for 5120 epochs. We stored the intermediate weights of each network at 10 different intervals ranging over 5 to 2560 epochs, training a total of 550 networks in the second phase.

Results

Experiment 1: Fixed RMSE Criterion

Pretraining

Figure 3 shows the number of training epochs for the two phases averaged across the 20 networks for each pretraining condition. The Non-experts required a much shorter training period than all the expert networks for the pretraining phase, reflecting the ease of basic level classification. For the second phase, the Non-experts were significantly slower than all the Experts in learning the new task ($p < 0.001$, pairwise comparison between Non-experts and the face, can, cup, book experts with $t(38) = 7.03, 5.74, 14.69, 10.76$, respectively). The difference between the can experts and the face experts was insignificant ($t(38) = 1.20, p > 0.2$), the face experts

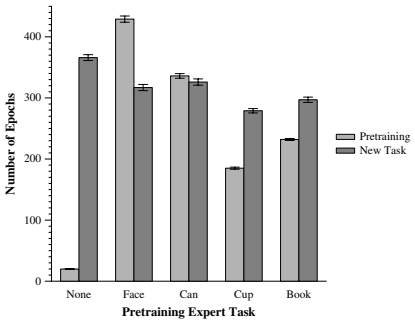


Figure 3: Number of epochs to reach the RMSE threshold. Error bars denote standard error.

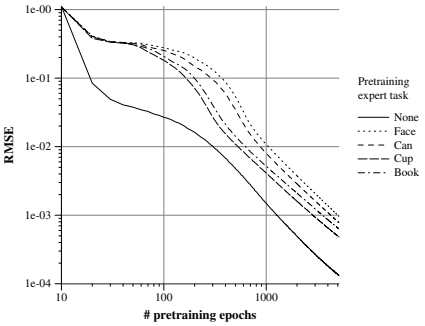


Figure 4: Learning curve for the pretraining phase

Table 1: Training set accuracy for just the Greebles. Figures in parentheses denote standard error.

Expert task	Greebles training set accuracy(%)
Non-expert	71.2 (2.00)
Face	93.5 (1.17)
Can	95.2 (1.04)
Cup	89.8 (2.00)
Book	88.8 (1.46)

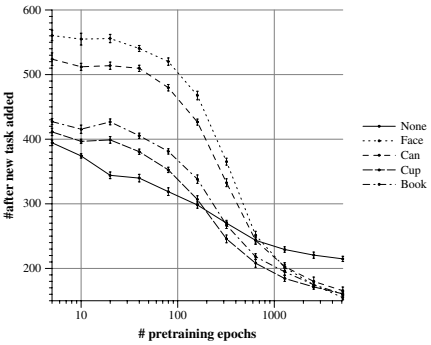


Figure 5: Number of epochs to learn the new task. Error bars denote standard error.

were slower than the book experts ($t(38) = 3.08, p < 0.005$), and the book experts were slower than the cup experts ($t(38) = 3.22, p < 0.005$).

Table 1 shows that despite the overall RMSE having been controlled, the Non-experts were still non-experts at Greebles after training on them. Further training on the Non-experts would have widened the gap between training times on the Greebles for Experts and Non-experts even more. On the other hand, for the Experts, there was a positive correlation between the training set accuracy and the number of training epochs, suggesting the differences in training epochs between the Experts would have narrowed if the training set accuracy on the newly added task had been controlled. Being an expert on some task prior to learning another expert task was clearly advantageous.

Experiment 2: Fixed Exposure Pretraining

Not surprisingly, the Non-experts maintained lower RMSE than all the Experts during the pretraining phase (Figure 4). Among the four groups of expert networks, the face experts had the most difficult pretraining task, followed by the can experts, book experts, and finally cup experts.

For the secondary task training, there was a crossover in effects at 1280 epochs: Fewer epochs meant the non-Experts had an advantage; more meant Experts had an advantage (Figure 5). If the networks were pretrained long enough, the improvement on the error for the pretrained portion of the task became negligible compared to the error due to the newly added task. In this case, we can safely argue that the epochs of training required in the second phase are fully determined by the learnability of the newly added task. If the pretraining stopped prematurely, however, the networks must improve their performance on the prior task as well as the newly added task to achieve the second phase RMSE threshold. All of the networks that were pretrained for at least 1280 epochs achieved a pretraining RMSE that was an order of magnitude lower than the second phase RMSE threshold. Therefore, the advantage that Experts with this amount of pretraining gained must be due solely to their performance in learning the new task.

Analysis: Network Plasticity We hypothesized that the advantage of learning a fine-level discrimination task would be due to greater plasticity in the hidden units. That is, we expected the activations of the hidden units to end up in the linear range of the squashing function in order to make the fine discriminations. This is also a good place to be for back propagation learning, as the higher the slope of the activation function, the faster learning occurs. We therefore analyzed how the features extracted at the hidden layer were tuned by measuring the plasticity (average slope) of the pretrained networks. Our findings surprised us.

We defined a network's plasticity as the value of the derivative of the activation function at the activation level averaged across all hidden layer units and all patterns in

a given set of input patterns:

$$P(S) = \frac{1}{N}\sum_{s\in S}\frac{1}{n}\sum_{i\in I} g'(x_{si})$$

where $g(x)$ is the activation function, S a set of patterns, N the number of patterns in S, I the set of hidden layer units, n the number of hidden layer units, and x_{si} the activation of unit i in response to pattern s. In the online backpropagation learning rule, $g'(x)$ scales the weight changes of the hidden layer units with respect to the errors computed for the output layer. The plasticity of neural networks is usually taken to be predictive of the ability to learn new tasks (Ellis & Lambon Ralph, 2000).

As the activation function, all the hidden layer units in our neural networks used the logistic sigmoid function:

$$g(x) = \frac{1}{1+\exp(-x)}$$

where x is the weighted sum of the inputs, or the inner product of the input vector $\vec{z}$ and the weight vector $\vec{w}$:

$$x \equiv \vec{z}\cdot\vec{w}.$$

The first derivative of $g(x)$ can be written as

$$g'(x) = \frac{\exp(-x)}{(1+\exp(-x))^2} = g(x)(1-g(x)).$$

For $x \in (\infty, -\infty)$, $g(x)$ ranges over $(0,1)$ and $g'(x)$ over $(0, 0.25]$. $g'(x)$ is a bell-shaped function with a global maximum at $g'(0) = 0.25$. By our definition of plasticity, the networks that produce intermediate responses at the hidden layer level would have higher plasticity than networks with bimodal responses. Networks with higher plasticity are generally more likely to learn new tasks faster since the hidden layer units would change their weights more rapidly in response to the errors propagated from the newly added output units.

Network plasticity can also be considered as a measurement of *mismatch* between the hidden layer weights and the input patterns. If the input patterns and the weights were orthogonal, x would be near 0, resulting in maximal plasticity. If, however, the weights tuned for some task matched the input patterns of a new stimulus class, $|x|$ would have a larger value resulting in lower plasticity. The issue is whether these features will be advantageous for learning the new stimulus class. When a network with a low plasticity (high match) measured on novel patterns learns faster on those patterns than a network with higher plasticity, this suggests that the highly matched features are efficacious for classifying the new stimuli.

Figure 6 shows the plasticity of the pretrained networks in response to the training set used for the pretraining and to the set of new patterns which would be added into the training set for the second phase. For both the pretrained patterns and the unseen patterns, Non-experts

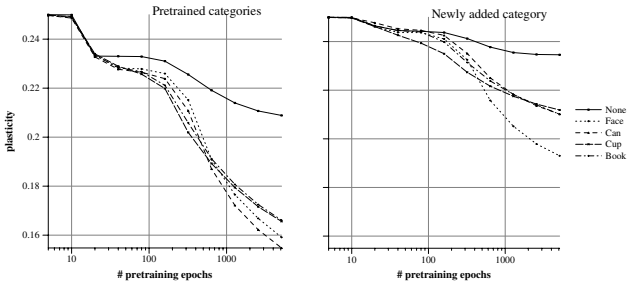


Figure 6: Plasticity of the pretrained networks

retained their plasticity better than all Experts. As we saw in the previous section, however, it was the Experts that eventually showed an advantage in learning the new task. All Experts learned the new task faster as they rapidly lost plasticity over pretraining time. Normally, we would expect the Experts to be generally poorer in learning new tasks due to their low plasticity. These results imply that the advantage the Experts gained in learning the Greebles task cannot be explained as part of a general capability of learning new tasks. Instead, it is more appropriate to interpret this to mean that the hidden unit features were well-matched to the new task.

The lower plasticity of the Experts for the pretrained data implies that the Experts had to finely tune their features to fit their respective expert category, while the Non-experts did not require fine tuning of the features to achieve the lower errors. The eventual advantage of the Experts can then be explained in terms of the features tuned for the expert tasks matching the Greebles data set as well. Given the strong trend regardless of the domain of expertise, we claim that the features useful for one subordinate classification task are general expert features that are good for discriminating individuals of other classes as well. Although other uncontrolled factors such as the length of the weight vectors can influence the plasticity of a network, they seem unlikely to explain our experimental results.

Experiment 2 Summary For longer pretraining epochs, the Non-experts took longer than any of the Experts to reach the final RMSE after the new task was added. While we would expect the Experts to have higher plasticity given their advantage in learning the new task, it was the Non-experts that retained higher plasticity. A comparison between Experts within each task also showed that the networks with longer pretraining and lower plasticity learned the new task faster. The results regarding network plasticity led us to interpret plasticity as a measurement of mismatch specific to a given set of patterns, rather than a predictor of the ability to learn arbitrary new tasks. Given these results, we claim that the underlying cause for the advantage gained by the Experts is the generality of the hidden layer features, fitting well with the subordinate classification task of other classes. This is remarkable in that *overtraining* on a prior task facilitates learning the new one.

Conclusion

Based on the recent studies that showed FFA's engagement in visual expertise of homogeneous stimuli is not limited to faces, we hypothesized that the features useful for discriminating individual faces are useful for the expert learning of other classes. Both of our experiments yielded results in favor of our hypothesis. Furthermore, while faces had a tendency to show the greatest advantage, the results were replicated with networks whose expertise was with other stimulus classes, including cups, cans and books.

The results of the two experiments showed that the possession of a fully developed expertise for faces or non-face objects is advantageous in learning the subordinate level classification of Greebles. Contrary to our expectation that expert networks would show greater plasticity, analyses of network plasticity for Experiment 2 showed that plasticity decreased for the expert networks over time, and it was lower than for non-Expert networks. Indeed, the *lower* the plasticity, the *less* time it took to learn the new task. By reinterpreting low plasticity to mean "high match," we take these results to mean that the features being learned not only match the Greeble stimuli well, but also are the *right* features for fine discrimination of Greebles. Since the choice of Greebles for the second experiment was arbitrary, this suggests that learning to discriminate one homogeneous visual class leads to faster learning in discriminating a new one. Therefore, we conclude that visual expertise is a general skill that translates well across a wide variety of object categories.

Future Work

Firm believers in the face specificity of the FFA might insist that it must be shown that individual neurons in the FFA can simultaneously code features for multiple classes of objects in order their theory to be rejected (Kanwisher, 2000). Even with the advances in brain imaging technology, monitoring each neuron in the FFA is infeasible. Simulations with computational models, however, allow us to monitor the behavior of every single unit in the network.

Naturally, then, one possible extension of the current research is to investigate in detail what the hidden layer units in the expert networks are encoding. Although our experimental results seem to suggest there are features that are useful for visual expertise of any object classes, it is unclear exactly what those features are. Visualization of these expert features would help understand how we develop visual expertise.

References

Buhmann, J., Lades, M., and von der Malsburg, C. (1990). Size and distortion invariant object recognition by hierarchical graph matching. In *Proceedings of the IJCNN International Joint Conference on Neural Networks*, volume 2, pages 411–416, New York. IEEE.

Dailey, M. N. and Cottrell, G. W. (1999). Organization of face and object recognition in modular neural network models. *Neural Networks*, 12(7–8):1053–1074.

De Renzi, E., Perani, D., Carlesimo, G., Silveri, M., and Fazio, F. (1994). Prosopagnosia can be associated with damage confined to the right hemisphere — An MRI and PET study and a review of the literature. *Psychologia*, 32(8):893–902.

Ellis, A. and Lambon Ralph, M. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 26(5).

Farah, M. J., Levinson, K. L., and Klein, K. L. (1995). Face perception and within-category discrimination in prosopagnosia. *Neuropsychologia*, 33(6):661–674.

Gauthier, I., Behrmann, M., and Tarr, M. J. (1999a). Can face recognition really be dissociated from object recognition? *Journal of Cognitive Neuroscience*, 11:349–370.

Gauthier, I., Skudlarski, P., Gore, J. C., and Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, 3(2):191–197.

Gauthier, I. and Tarr, M. (1997). Becoming a "greeble" expert: Exploring mechanisms for face recognition. *Vision Research*, 37(12):1673–1682.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., and Gore, J. C. (1999b). Activation of the middle fusiform "face area" increases with expertise in recognizing novel objects. *Nature Neuroscience*, 2(6):568–573.

Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience*, 3(8):759–762.

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17:4302–4311.

Lades, M., Vorbrüggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Würtz, R. P., and Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311.

Moscovitch, M., Winocur, G., and Behrmann, M. (1997). What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *Journal of Cognitive Neuroscience*, 9(5):555–604.

The Role of Feedback in Categorisation

Mark Suret (m.suret@psychol.cam.ac.uk)
Department of Experimental Psychology; Downing Street
Cambridge, CB2 3EB UK

I.P.L. McLaren (iplm2@cus.cam.ac.uk)
Department of Experimental Psychology; Downing Street
Cambridge, CB2 3EB UK

Abstract

The most popular theories of associative learning require some kind of external teaching signal, whether it be feedback on category membership or the presence of a US to determine whether a trial has been successful. However, under some circumstances (Bersted, Brown & Evans, 1969; Evans & Arnoult, 1967; Estes, 1994), it appears that such a signal is not required to produce accurate categorical decisions. The work presented here was motivated by Wills & McLaren (1998), which looked at Free Classification and demonstrated that human participants could accurately extract category structure with no feedback during the task. The current experiments compare performance on a generalisation task after training using one of a number of different conditions including Free Classification, allowing a comparison with participants who have had a teaching signal during training. The results seem to indicate that under some conditions, it is not feedback during training, rather the need for a decision to be made which is critical. The data also shows that there is no difference in accuracy between the ability of the groups to generalise, no matter what their initial training, rather the differences are manifested in the reactions times.

Introduction

There has been little work done on directly comparing performance on tasks where one group of participants receives feedback and another does not, as it seems intuitively obvious that if there is some external feedback, then this will aid rather than retard learning. The familiar notion that consistent feedback always produces the best performance (Homa & Cultice, 1984) can be challenged, however, with data showing that a teaching signal may not provide any advantage. Real world scenarios, where feedback may be absent, and theories of natural categorisation (Rosch & Mervis, 1975) seem to be in line with this idea. If we allow that reliable feedback is not always the key to success, then there must be other factors which participants are only able to take advantage of in some circumstances. Estes (1994) provides an experiment where two different types of feedback are compared. Observational training is the same as the Label condition described later in this report, where a label is provided with the stimulus and denotes which category it belongs to, with no decision being made by the participant.

This is compared to standard training, which has corrective feedback after a decision has been made about the category membership of a given stimulus by the participant. In Estes' study, the observational training was found to be more effective than the standard training on test, and it was deduced that the observational training was more consistent with the participants' perception of the categories. Both these types of training were included in the current experiment, as well as a free classification condition, and two other conditions designed to control for potential effects of response practice.

Experiment 1

Participants and Apparatus

The participants were 60 adults, aged between 18 and 35, who were paid for their participation. All were graduate or undergraduate students from the Cambridge area. Participants were tested individually in the same quiet experimental cubicle. The room contained an Acorn RISC PC600 microcomputer connected to a 14 inch colour monitor, model AKF60. Responses were made using a standard keyboard. Participants sat about one metre away from the screen which was approximately at eye level.

Stimuli

Each stimulus was a 16 × 16 array of black and white squares. These "chequerboards" measured 2.5 cm on a side and were presented on the centre left of the screen against a mid grey background. In some experimental conditions, a label for the chequerboard, either the letter 'A' or 'B' was presented in white on the centre right of the screen. The label was approximately the same size as the stimulus. For each participant, a new master pattern was created, and this was used as one of the prototypes during the experiment. The master pattern was a chequerboard of 128 white and 128 black squares randomly arranged, hereafter chequerboard A. A second prototype was created from the master pattern by selecting 120 of the squares (60 black and 60 white) and reversing the shade of those squares (black to white or vice versa), hereafter chequerboard B. The examples that were actually shown to the participants depended upon the phase of the experiment. During training the stimuli shown were created from the prototypes by subjecting each square to a small independent chance ($p = 0.05$) of reversing its colour. On test all stimuli were created from the master pattern and had a set number of squares reversed in colour, ranging from 0 to 120 in steps of 10, covering the artificial continuum between the prototypes. Each test stimulus reversed the colour of a different, randomly selected set of squares, depending on its position along the continuum between the prototypes. For example if a stimulus was in

position 4 out of 12, then 40 of the master pattern squares would be changed to be the same as those in the second prototype. Another stimulus at position 4 would have a different set of 40 squares reversed in colour so as to match those in the other prototype.

Design

The experiment incorporated one of five different training conditions, all followed by an identical test session.

Labelled

Participants were presented with the training chequerboards accompanied by a consistent label, either A or B during training. No response was required and the participants were asked to try to learn what features made a pattern either an A or a B.

Free Classification

Participants were presented with the training chequerboards, but without a label being present. They were asked to divide the stimuli that they were being shown into two groups, in any way which they saw fit, by pressing one of the appropriate keys. No feedback was given throughout.

Corrective Feedback

Participants were presented with the training chequerboards alone and had to learn which of the two keys to press when they saw a given chequerboard. Corrective feedback in the form of a beep from the computer indicated when the participant had pressed the wrong key. No label was presented.

Mental Decision

Participants were given similar instructions to the Free Classification group, except no response was required in this condition. All the participants had to do was decide to which one of the two groups the pattern belonged. No label was presented.

Matching to Label

Participants were given similar instructions to the Labelled condition, except that they were required to make a keypress once the chequerboard had disappeared. The key they pressed simply corresponded to the label that was presented with the stimulus. This same key assignment was carried over into the generalisation phase.

Procedure

Each participant was asked to read the general experimental description which was displayed on the computer screen. This included a brief description of the whole experiment, more detailed instructions about the training phase and an example stimulus, with a label if appropriate. Participants were told to try to learn as much as they could in the first part of the experiment, as they would need this information for the final part. Once the participant had read the instructions, the experimenter verified that they had understood them, and then left the room for the remainder of the experiment. Participants started the training phase by pressing the 'Y' key at the top of the keyboard, and then proceeded through the 60 training trials in a fashion determined by the experimental condition.

During training, the stimuli (and label if appropriate) were displayed for 5 seconds before disappearing from the screen. If a response was required, then participants were allowed to press an appropriate key, either the 'x' or the '>' key, once the stimulus had gone from the screen. If no response was required during training, then there was a two second inter-stimulus interval. Training consisted of 30 presentations of A and 30 of B in a random order. All participants were told that the chequerboards that they were to be shown could be divided into two groups, and that this was their task.

After the final training trial, the instructions concerning the test phase were displayed. The test phase was identical for all participants regardless of their training condition and consisted of 130 trials which displayed a chequerboard on its own. The instructions asked each participant to continue placing the stimuli into two groups in the same way as before, but were informed that there would now be no label, if there had been one before, and that a response was required for each chequerboard. In conditions where a label had been presented to the participants a key mapping was provided, for example, "Press 'x' if it's an A". If no label had been present during training, participants were simply asked to carry on placing the chequerboards into the most relevant group. The instructions for the Mental Decision group required them to make an arbitrary assignment of the groups they had formed during training to the keys to be used during the test phase. Once a response to a test stimulus had been made, it was immediately replaced by another stimulus. If any other key apart from the two that were designated was pressed, the computer beeped, and another response was required. Participants were asked to focus more on accuracy rather than speed in this part of the experiment, and there was no explicit time-out procedure, so participants cannot be considered to be under any time pressure.

The design of the test phase was such that participants were shown stimuli along the continuum from the master pattern (A) to the second prototype (B). Each test stimulus had a multiple of 10 squares changed from A, to make it more like B. As A and B differed by 120 squares, there are 13 steps along such a continuum and with each point being sampled 10 times, this gives 130 test trials. The 10 stimuli for any given point on the continuum were all different, and generated as described above.

At the end of the experiment the computer automatically recorded responses and reaction times from the training phase where possible, and from the test phase in a data file. Each participant was paid for their time and thanked for their participation.

Results

The two measures of performance recorded during the generalisation test phase were response and reaction time, and they are dealt with separately. The independent measure on the plots in this section is distance along the

Table 1: Factors in Experiment 1

	Information	No Information
Keypress	Match to Label	Free Classification
No Keypress	Label	Mental Decision

B-A continuum, with B at one end and A the other, and the intermediate values being examples of B with (10 × distance) squares changed to make the stimulus more like A. Data was analysed firstly by comparison of the groups using ANOVA. A factorial design was also used, as illustrated below (Table 1). The conditions included in the experiment have tried to control for possible effects of making a response interacting with the presence of accurate information about category membership.

Responses

A single mixed design analysis of variance (ANOVA), with one within-subject variable (continuum position, 13 levels) and one between subjects variable (training condition, 5 levels) was performed on the mean number of B responses at each point on the continuum for each participant. This failed to reveal any significant difference between the groups $F(4,55) = 2.25$, $p<0.1$ or any interaction between the groups and the position along the continuum $F(48,660) = 0.647$, $p>0.9$. There was a significant main effect of position along the B-A continuum $F(12,660) = 6.88$, $p<0.001$. No further analyses were carried out on the response data, as there was no clear difference between the groups at this stage. The mean responses for each group at each point along the continuum are plotted in Figure 1. A mixed design ANOVA, with one within subject variable (continuum position, 13 levels) and two between subject variables (the presence or absence of consistent information about category membership and the requirement to make a keypress) was performed. The analysis yielded a significant main effect of both the requirement to make a keypress, $F(1,44) = 4.76$, $p < 0.05$ and of continuum position, $F(12,528) = 99.45$, $p < 0.001$. No other effects approached significance in this analysis, $p > 0.15$. The mean responses for each level at each point along the continuum are plotted in Figure 2.

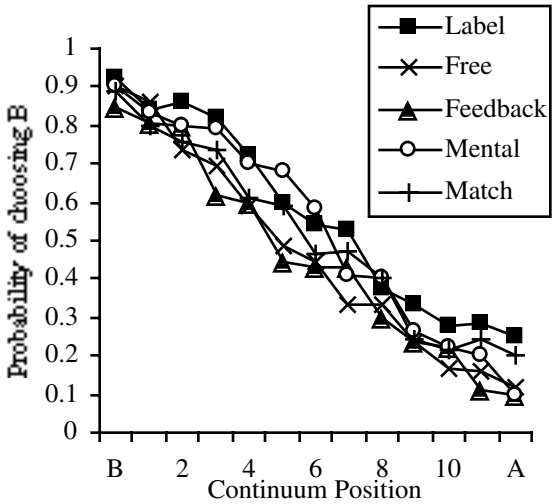


Figure 1: Response Data for All Groups in Experiment 1

Reaction Times

Participants were not considered to be under any time pressure, so the reaction times provide a secondary performance measure. The mean reaction times for each group are plotted against the distance from B along the B-A continuum in Figure 3. ANOVAs were performed on the mean reaction times for each participant. A single mixed design ANOVA, with one within subject variable (continuum position, 13 levels) and one between subjects variable (training, 5 levels) revealed a significant main effect of group, $F(4,55) = 5.00$, $p<0.01$, and of position along the continuum, $F(12,660) = 6.28$, $p<0.001$. There was no significant interaction between the two factors, $F(48,660) = 1.18$, $p > 0.15$.

A Tukey HSD test performed on the group factor revealed that each of the Free Classification, Mental Decision and Feedback conditions were significantly faster than the Label condition, all $p< 0.05$, with no other comparisons reaching significance. Pairwise analysis of the conditions reveals a significant quadratic interaction between the Label and Free Classification conditions, $F(1,22) = 5.84$, $p < 0.05$, and the Matching to Label and Free Classification conditions, $F(1,22) = 7.05$, $p < 0.05$, no other comparison of quadratic trends reached significance, $p > 0.1$.

A mixed design ANOVA, with one within subject variable (continuum position, 13 levels) and two between subject variables (the presence or absence of consistent information about category membership and the require-

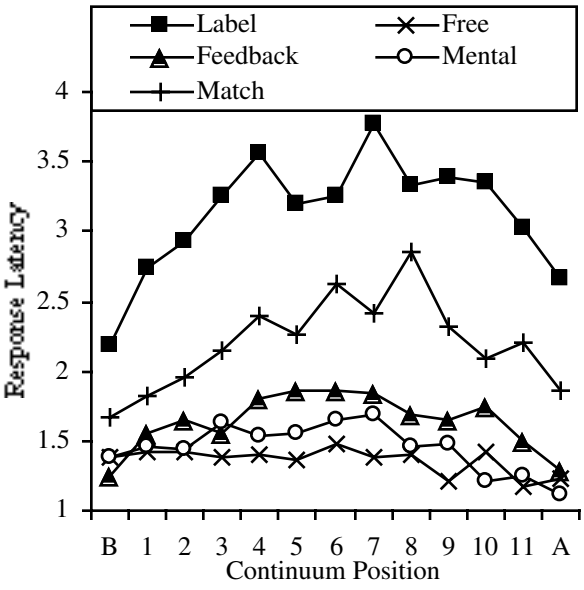


Figure 3: Reaction Times for All Groups in Experiment 1

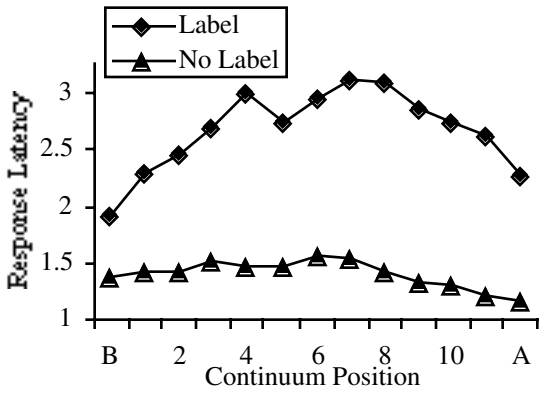


Figure 4: Effect of Label in Experiment 1

ment to make a keypress) was performed. This revealed a significant effect of information, $F(1,44) = 13.67$, $p = 0.001$, the expected main effect of continuum position, $F(12,528) = 4.83$, $p < 0.001$ and an interaction between these two factors $F(12,528) = 2.77$, $p < 0.05$. No other effects approached significance in this analysis, $p > 0.25$. The mean reaction times for each level at each point along the continuum are plotted in Figure 4.

Discussion

Experiment 1 provides evidence for a number of novel results with respect to the effect of feedback on categorisation. Whilst there is no significant observable effect of training condition on response performance, there is a significant effect on the reaction time data collected. Initially surprising is the fact that groups trained with a completely errorless teaching signal (Labelled and Match to Label) recorded the slowest reaction times. Two other groups received no feedback (Free Classification and Mental Decision), and both are found to be significantly faster than the Labelled condition. The response curves show that this speed advantage cannot be due to speed-accuracy trade off, and so there must be some other explanation for the deficit in performance observed in groups which intuitively should be the best at the task.

The main effect of labelling in the factorial analysis for the reaction time data confirms this finding, with the presence of a category label causing an increase in response time on generalisation. The effect of making a keypress can be seen in the response analysis, but can be regarded as relatively uninteresting as there is no sign of an interaction which might point to the task having been learnt better, causing one group to have a more step-like function.

The most obvious reason for the speed advantage seen for those groups who do not have a consistent piece of information relating to category membership is that they are required to make a decision about category membership during the training phase. This seemingly essential part of training is not present when a label is provided, as the stimulus and its category name are presented simultaneously. In the Feedback condition, even though there is consistent category information, a decision must be made before this feedback can be received and integrated. In the conditions where no feedback it present, a decision made internally about previous stimuli is the only information available when deciding to which group subsequent stimuli should be assigned. The similarity of the generalisation functions implies that all groups have learnt how to differentiate between the two categories to the same extent. However it may be that the application of this knowledge is mediated by a response mechanism which is yet to be set up by those groups which are not required to make a decision as to category membership during training. This leads to the difference seen between the relatively flat reaction time curves for the three conditions where decisions are required in training and the inverted U-shaped reaction time curves produced by those participants who are presented with the label during training. The inverted U-shaped curves are typical of those produced during generalisation along a continuum using these procedures (Jones, Wills & McLaren, 1998). The centre point of the continuum is no more like an A than a B, so any response made to these stimuli must be indeterminate, and hence produces a longer latency than those responses to items which are more like those in training.

Despite the potentially simple explanations for the differences observed between the groups, it is still an interesting result to have the Free Classification and Feedback conditions indistinguishable from one another even with the supposed added advantage of corrective feedback. This may be due to some motivational factor, as participants in the Free Classification group are never told that they are wrong, however with feedback, participants may be relatively sure that the stimulus that they are seeing is an A, but in fact turns out to be a B, which may disrupt their representation of the conditions for category membership. The feedback that they get may be consistent within the framework of the experiment, but may be inconsistent internally, and this may be part of the reason for the lack of benefit for the Feedback condition.

The reason the Label condition is so slow may be because it takes time to learn, and then to use, the response mapping. It cannot be entirely due to motor learning as although there is some advantage for the Matching to Label group over the Label group, there is still a difference between the Matching condition and the three where a decision was made during training. The results from the Mental Decision group also tend to discount this line of reasoning as they were not required to form a key mapping before the test phase but their responses are indistinguishable from the Feedback and Free Classification.

Experiment 2

Experiment 2 was designed in a factorial fashion to investigate the effect of both feedback and consistent label-

Table 2: Factorial Design of Experiment 2

	Feedback	No Feedback
Label	Match to Label (FB)	Match to Label (NFB)
No Label	Corrective Feedback	Free Classification

ling of the stimuli. In this respect it was similar to Experiment 1, but attempted to control for the formation of response mappings during training. Experiment 2 forced all participants to adopt a key mapping during training, so that any effect on test would be due to the actual training rather than differences in procedure. All participants were forced to make their responses to the stimuli within two seconds of the stimulus disappearing using the same keys as before.

Stimuli and Apparatus

These were identical to those in Experiment 1.

Participants and Design

Forty-eight Cambridge University students took part in the experiment. All were aged between 18 and 25. The experiment was designed to test the effect of two factors when learning an artificial categorisation problem. These were the presence and absence of feedback on the responses that were made during training, and the presence or absence of a consistent category label during training The design was similar to Experiment 1, with the test phase being identical, and is shown in Table 2.

Procedure

The procedure was similar to Experiment 1. The only differences were during training. All groups were presented with a stimulus and asked to respond within two seconds to that stimulus once it had disappeared. The training differed from Experiment 1 by presenting the label, if necessary, before the stimulus rather than concurrently. For those conditions without a label, a '#' symbol was presented before each stimulus, whether it was nominally an A or a B, in place of the label to equate the training time. The appropriate label or '#' was displayed for five seconds before the five second presentation of the stimulus. The instructions were identical to those from Experiment 1 apart from detailing the separate presentation of the label or '#' and the stimulus and informing the participants of their two second time limit.

Results

As with Experiment 1, there were two dependent variables, response selection and reaction time, and as a time limit had been imposed, the reaction times can be considered a more informative performance indicator. The data were analysed using an appropriate ANOVA. The performance of individual groups was analysed along with the effect of the factors built into the experiment.

Responses

A single mixed design analysis of variance (ANOVA), with one within-subject variable (continuum position, 13 levels) and one between subjects variable (training condition, 4 levels) was performed on the mean number of A responses at each point on the continuum for each participant. This failed to reveal any significant difference between the groups $F(3,44) = 0.76$, $p>0.5$ or any interaction between the groups and the position along the continuum

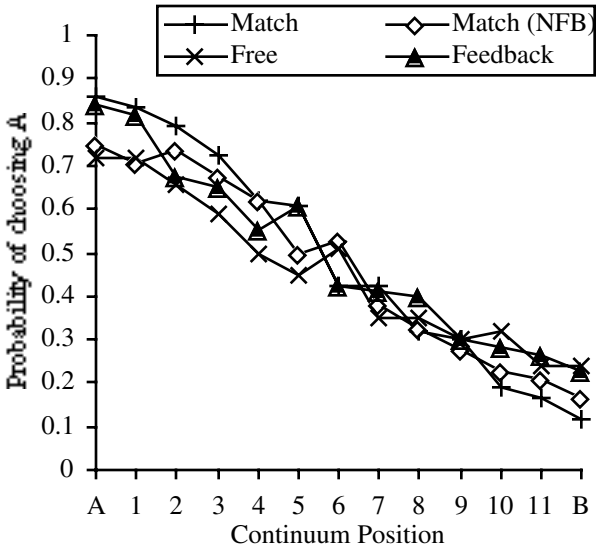


Figure 5: Response Data for All Groups in Experiment 2

$F(36,528) = 0.65$, $p>0.5$. There was a significant main effect of position along the A-B continuum $F(12,528) = 55.73$, $p<0.001$. No further analyses were carried out on the response data, as there was no clear difference between the groups at this stage. The mean responses for each group at each point along the continuum are plotted in Figure 5.

A mixed design ANOVA, with one within subject variable (continuum position, 13 levels) and two between subject variables (the presence or absence of consistent information about category membership presence of feedback) was performed. The analysis yielded only a significant main effect of continuum position, $F(12,528) = 55.73$, $p < 0.001$. No other effects approached significance in this analysis, $p > 0.15$.

Reaction Times

The mean reaction times for each group are plotted against the distance from A along the A-B continuum in Figure 6. ANOVAs were performed on the mean reaction times for each participant.

A single mixed design ANOVA, with one within subject variable (continuum position, 13 levels) and one between subjects variable (training, 4 levels) revealed a significant main effect of group, $F(3,44) = 3.38$, $p = 0.026$, and of position along the continuum, $F(12,528) = 5.64$, $p < 0.001$. There was no significant interaction between the two factors, $F(36,528) = .90$, $p > 0.6$. A Tukey HSD test performed on the group factor revealed that the Match to Label condition with feedback was found to be significantly slower than the Free Classification condition, $p < 0.05$. A mixed design ANOVA, with one within subject variable (continuum position, 13 levels) and two between subject variables (the presence or absence of category membership information and the presence of feedback) was performed. This revealed a significant effect of category information, $F(1,44) = 4.54$, $p = 0.039$, with the label conditions showing the longer mean

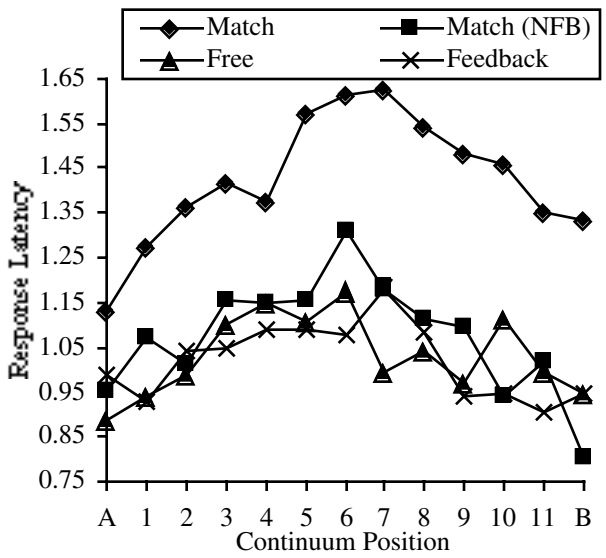


Figure 6: Reaction Time Data for All Groups in Experiment 2

reaction times, and the expected main effect of continuum position, $F(12,528) = 5.64$, $p < 0.001$. No other effects were significant in this analysis, $p > 0.09$.

General Discussion

The results from Experiment 2 are broadly in line with those obtained in Experiment 1. The inclusion of a consistent category label appears to have a detrimental effect when compared with the requirement to make an active decision about category membership. However, there appears to be a speed-accuracy trade off present in the results. Whilst the Match to Label with feedback (MFB) group are slowest, they also show the greatest difference between the ends of the generalisation gradients. Although this difference is not significant it would be difficult to conclude anything definite on the basis of this alone. However, taken with the results from Experiment 1, it seems clear that the fact that feedback is not present does not seem to have a detrimental effect on the performance shown by participants. Instead it seems that, in some cases, providing an entirely consistent label for the stimuli during training causes participants to perform worse. This is not what would be predicted from Homa & Cultice (1984), and is at odds with the results from Estes (1994) who showed that a condition analogous to the Label condition in Experiment 1 gave better performance on test than when participants were trained using a corrective feedback approach. It may be that the real advantage lies in being able to make active (i.e. self-generated) decisions during training rather than simply being exposed to the stimuli and the appropriate category information (Figure 4). Thus it may be the case that different processes are at work

Conclusion

It seems likely that the most successful approach to modelling such data will come form simple self-organising systems (Rumelhart & Zipser, 1986; Saksida, 1999) which are able to extract the necessary information from the stimuli encountered to form coherent categories through exposure to the stimuli alone. It may be the case that all that is needed is exposure to stimuli in order to extract information about them, and this raises the interesting question of what exactly feedback does if it does not always aid decision making.

References

Bersted, C.T., Brown, B.R. & Evans, S.H. (1969). Free sorting with stimuli in a multidimensional attribute space. *Perception and Psychophysics, 6B,* 409-413.

Estes, W.K. (1994). *Classification and Cognition.* Oxford: Oxford University Press.

Evans, S.H. & Arnoult, M.D. (1967). Schematic concept formation: Demonstration in a free sorting task. *Psychonomic Science, 9(4)*, 221-222.

Homa, D., & Cultice, J. (1984). Role of Feedback, Category Size, and Stimulus Distortion on the Acquisition and Utilization of Ill-Defined Categories. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10,* 83-94.

Jones, F.W., Wills, A.J. & McLaren, I.P.L. (1998). Perceptual Categorisation: Connectionist modelling and decision rules. *Quarterly Journal of Experimental Psychology, 51(B),* 33-58.

Rosch, E., & Mervis, C.B. (1975). Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology, 7,* 573-605.

Rumelhart, D.E., & Zipser, D. (1986). Feature discovery by competitive learning. In D.E. Rumelhart, J.L. McClelland, & The PDP research group. *Parallel Distributed Processing: Explorations in the microstructure of cognition.* Cambridge, MA: MIT Press.

Saksida, L.M.(1999). Effects of similarity and experience on discrimination learning: A nonassociative connectionist model of perceptual learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *25*, 308-323.

Wills, A.J. & McLaren, I.P.L. (1998). Perceptual Learning and Free Classification. *Quarterly Journal of Experimental Psychology, 51(B),* 235-270.

An Analogue of The Phillips Effect

Mark Suret (m.suret@psychol.cam.ac.uk)
Department of Experimental Psychology; Downing Street
Cambridge, CB2 3EB. UK

I.P.L. McLaren (iplm2@cus.cam.ac.uk)
Department of Experimental Psychology; Downing Street
Cambridge, CB2 3EB. UK

Abstract

Previous experimental work has demonstrated that human participants can easily detect a small change in a visual stimulus if no mask intervenes between the original stimulus and the changed version and the inter-stimulus interval is small (Phillips, 1974). Rensink, O'Regan & Clark (1997) have shown that if a mask is used then detecting the change is extremely difficult, no matter how small the ISI is made. This work attempts to establish whether familiarity with a stimulus has any effect on a participants ability to detect a small change in it using Rensink's masking procedure with Phillips' stimuli (checkerboards). Participants were required to make judgements as to whether two stimuli, which alternated with one another in presentation, were the same or different. Some participants attempted the task using just one checkerboard pattern which became increasingly familiar across sessions, others were given new, randomly generated checkerboards for each trial. In both conditions, any change (which would occur on 50% of trials) would only affect one square of the pattern. The results show a clear advantage for the participants dealing with familiar stimuli in detecting any change, and go some way towards explaining why this is so.

Introduction

Phillips (1974) demonstrated how easy it was for participants to detect a change between two stimuli if they were presented one after the other without a gap in a single alternation. This is the Phillips Effect. He also investigated the consequences of inserting a grey mask and a blank screen between the two stimuli. The inclusion of an inter-stimulus interval adversely affected participants' performance, and the presence of a mask made performance even worse. Rensink et al. (1997) demonstrated that the brief inclusion of a grey mask between repeated presentations of two slightly different stimuli made any change extremely difficult to detect. This is the Rensink Effect. Their experiment used electronically altered images which allowed manipulation of the colour, position and presence of an object. Without the mask, spotting the difference becomes trivial, if the stimuli are positioned in the same place and then alternated. Current explanations of this phenomenon cite retinal transients (Klein, Kingstone & Pontefract, 1992) as the mechanism for detecting changes in this latter case, which would be unaffected by familiarity.

Some pilot work using a single participant indicated that the effect of familiarity with the stimuli was likely to be very significant. Hence introducing the notion of familiarity removes some of the difficulties intrinsic to the Rensink Effect and makes the task more similar in difficulty to the Phillips Effect. One drawback of this pilot experiment was that it contained familiar and random trials in each session so after a while the participant became able to tell which were the familiar trials, and this may have differentially affected the responses to each trial.

Nevertheless, the results of the pilot experiment (Figure 1) allowed the prediction that the Familiar condition would lead to better detection of changes than the Random condition. It was also predicted that there would be an improvement in performance as the amount of time spent on the task increased.

The main aim of this work was to take the Rensink Effect, and attempt to ameliorate it, by allowing participants to practice on the same stimulus all the time. This would require the use of a different type of stimulus, as the repeated use of a real life scene would be impossible to control properly, so checkerboards were used as the training stimuli. In this way, the participants were presented with essentially the same stimulus on every trial, but with the possibility of a change in one of the elements within the

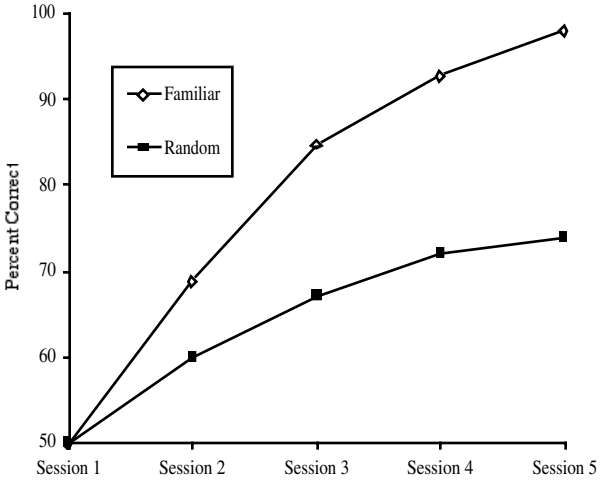


Figure 1: The Basic Familiarity Effect

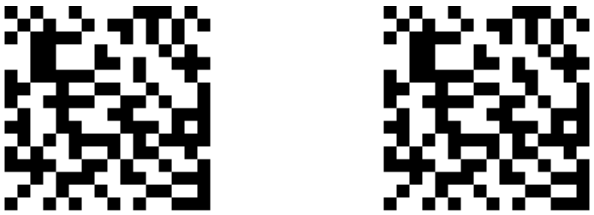


Figure 2: An example pair of chequerboard stimuli.

checkerboard. An example pair is shown in Figure 2) with the difference being rather difficult to spot.

Once a reliable difference had been found between the Familiar and Random groups, subsequent testing concentrated on the mechanism being used by participants to detect the changes.

Experiment

The experiment was conducted with two groups, each with different sets of stimuli. The Random group was a control group, and received different, randomly generated checkerboards for each trial during the experiment. In the Familiar group, each participant was trained on a single checkerboard unique to that participant. The primary aim of the experiment was to determine whether familiarity with the stimulus affected the participants' performance.

Initially two participants were run in each condition followed by a preliminary analysis. It was found that a large difference between the groups had already been established, and the final four participants were all run in the familiar condition. This allowed manipulations in the familiar condition, namely test blocks which departed from the standard task. The test blocks used manipulations intended to disrupt the performance of the participants in the Familiar group in the hope that this would suggest a possible mechanism for the way the changes were being detected. Clearly an effect of familiarity or session needed to be established first, as there is no directly relevant and properly controlled research in this area.

Stimuli and Apparatus

All the stimuli were randomly generated, two centimetre square checkerboards, with sixteen elements on a side giving a total of 256. Each base pattern stimulus had equal numbers of black and white squares, before any change was introduced. An example pair for the change condition is shown above (Figure 2), with the difference between the two checkerboards in the top right centre of the stimulus. Checkerboards were chosen as they are easy to manipulate for this type of experiment. Many different individual changes could be made whilst keeping the majority of the stimulus the same. In addition, the participants were unlikely to be familiar with the stimuli prior to the experiment

Those participants assigned to the random condition were given a newly generated checkerboard on each trial, whereas those in the familiar condition were always presented with the same pattern, albeit with a change on half the trials.

The experiment was run in a quiet room on an Apple Macintosh LCIII computer using a colour monitor. Participants responded to the stimuli by pressing one of two keys, either [x] or [.] on a QWERTY keyboard. Between blocks, participants were required to fill in a sheet to record their errors and reaction times before pressing the space bar to continue to the next block. The responses for each block were logged in separate data files.

Participants and Design

In total eight Cambridge undergraduates took part in the study. Four were allocated to the initial phase to determine the possible existence of a familiar/novel distinction. Two participants were allocated to the novel condition and two to the familiar condition. The remaining four participants were all allocated to the familiar condition.

The experiment consisted of a training phase for all participants and a test phase for those participants in the familiar group. The first four sessions were used for training for both groups, with the fifth session containing some test blocks for participants in the Familiar condition. All sessions for the random group were identical, as the tests given to the familiar group would have made no difference to a participant receiving a new checkerboard on every trial. Each one hour session consisted of ten blocks of stimuli, each containing 24 trials giving 240 trials in a session. Each block contained twelve trials where a change was present and twelve where there was no change between the two checkerboards. These trials were presented in a random order.

Participants were asked to try to detect a change between the two checkerboards, and respond appropriately as to whether or not they thought that a difference was present. For each trial, two checkerboards were alternated with one another, separated by a randomly generated mask. These checkerboards were either the same, or differed by one element within the pattern, i.e. one element that was black in one checkerboard was white in the other. The trials were such that each checkerboard was displayed for 500 milliseconds and the mask for 100 milliseconds. Over one trial, each checkerboard could be presented ten times, giving nineteen changes in a trial if the checkerboards were different. After the final alternation, the trial ended and if no decision had been made by the participant, then they were timed out.

During the fifth session, the participants in the familiar group were given test blocks in between familiar blocks, in an attempt to determine how they might be detecting the changes. These test blocks were ones containing random trials, such as those given to the participants assigned to the random group, and another type of block, labeled "C". In these blocks, there was always a fixed, random one square difference from the original base pattern, on both checkerboards, whether the trial was one of change or no change. This fixed change was different on each trial within the block. On change trials, there was also an additional change made to one of the checkerboards. This manipulation ensured that some difference from the base pattern was no longer a cue for change, although there was still a single change present between the two checkerboards

on change trials. The idea behind this manipulation was to contrast any changes in performance on "C" trials with that obtained on Random trials. In the former case, the perturbation of the familiar pattern is minimal, in the latter case it is, in some sense maximal. The sequence of blocks was: Familiar, "C", Familiar, Random, Familiar, "C", Familiar, Random, Familiar, "C". This gave three Familiar (as the first block is removed from any analysis), three "C" and two Random blocks to be used in the analysis for each participant from a session of ten blocks. The Familiar blocks were inserted between the test blocks to allow the participants an opportunity to re-establish baseline performance before the next test block was administered.

Procedure

The participants were seated in front of the computer approximately 50 centimetres from the screen and asked to make themselves comfortable. They were then read the instructions concerning their task, and were then asked if they had any questions about the instructions they had just been given. The participants were asked to respond to a "change" trial with their left index finger, by pressing the [x] key, and to a "no change" trial by pressing the [.] key with their right index finger. The participant was then asked to press the space bar to begin, and follow the on screen instructions that occurred throughout the experiment. The experimenter waited in the room until the first few trials had been completed, to ensure that the participant fully understood what it was that they were meant to be doing before leaving the room. Each block was started by pressing the space bar. The trials consisted of the alternation of two checkerboards, with a random black and white dot pattern mask being presented between presentations of the checkerboards. These checkerboards were either the same or differed by one element. The checkerboards subtended a visual angle of approximately two degrees and were presented in the centre of the screen. The participants were given feedback on each trial, with the words "correct" or "error" being displayed on the screen. If an error was made, the computer also beeped. After each block of twenty-four trials, participants were required to record their errors and reaction times on a sheet provided for them in the room. This was primarily to get the participants to take a break between blocks. It also gave a readily available source of data that could be tallied with the analysis on the computer. At the end of the session of ten blocks, the participants were given a short questionnaire to determine how motivated they were feeling during the session and what, if any strategy they were using.

After the questionnaire had been completed the participants were thanked for their time and the next session was arranged. After the fifth session, a more thorough questionnaire was administered, and the participants were paid and thanked for their participation.

Results

The basic familiarity effect is shown below in Figure 3. The Familiar and Random groups are denoted by F and R respectively. The graph shows that both groups improved at the task at roughly the same rate, but that performance in the Familiar group is better than that in the random group on all sessions by a roughly constant amount.

The initial analysis focused on finding significant effects of both group and session. Three dependent variables have been determined for each session: overall accuracy; percentage of changes detected; percentage of correct no change trials. Each of the three variables used may indicate something different about the way that the participants may be performing their task.

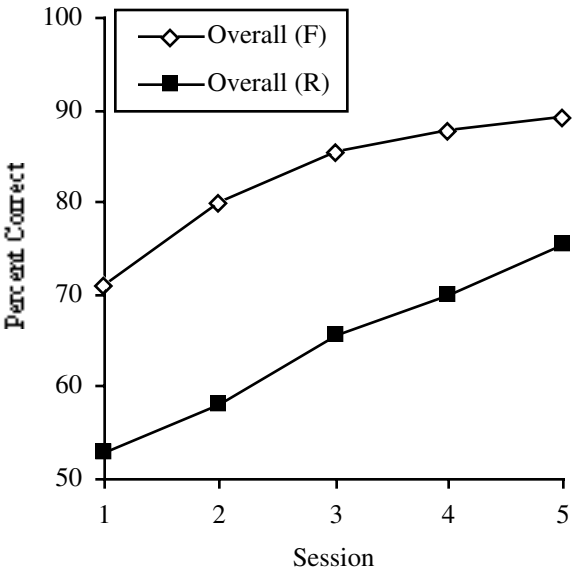


Figure 3: Between Groups Familiarity Effect.

Table 1: Within Session Comparisons between Random and Familiar Groups

(All probabilities are one-tailed)

	Overall Percentage Correct	Percentage of No Change Trials Correct	Percentage of Changes Detected
Session 1	U=0, p=0.022	U=0, p=0.023	U=0.5, p=0.032
Session 2	U=0, p=0.022	U=0, p=0.020	U=1, p=0.046
Session 3	U=0, p=0.022	U=0, p=0.022	U=0, p=0.022
Session 4	U=0, p=0.022	U=0, p=0.020	U=0, p=0.023
Session 5	U=0.5, p=0.033	U=0, p=0.017	U=1, p=0.048

Table 2: Between Session Comparisons Collapsed across Groups

(All probabilities are one-tailed)

Change between Sessions	Overall Percentage Correct	Percentage of No Change Trials Correct	Percentage of Changes Detected
1 and 2	T=1, p=0.009	T=5, p=0.034	T=2, p=0.013
2 and 3	T=0, p=0.009	T=2, p=0.036	T=0, p=0.014
3 and 4	T=2.5, p=0.024	T=6.5, p=0.100	T=6, p=0.046
4 and 5	T=6, p=0.173	T=0, p=0.008	T=8, p=0.156

Only non-parametric tests were used to analyse the data, as there were both small and unequal numbers in the groups. A significance level (α) of $p<0.05$ was used for all analyses. In each session, the score from the first block was not included in any subsequent analysis. This was to allow the participants an opportunity to practice the task before the session proper began. Means for each variable for a given session were used for all analyses. In the session 5 analyses, the different block types were separated and then the average for each variable was used.

Differences between Groups

This analysis compares the performance of the two groups on all three variables. For session five, only the familiar blocks are used for the analysis. The Mann-Whitney U-Test is used to compare the two unrelated samples.

The effect of group is so large that the difference between the Familiar and Random groups is significant for all variables in all sessions. In each case the percentage is higher for the Familiar group. A summary of the statistical results obtained from the analyses for comparisons between the Familiar and Random groups for each session is given in Table 1. Probabilities reported are one-tailed following the results of the pilot work discussed in the introduction.

Differences by Session

The test used in these analyses was the Wilcoxon Matched Pairs Test, as the values being compared were from the same participant, tracking their improvement as the sessions progressed (Table 2).

The effect of session is also a major factor in the performance of all the participants. Early in training there is the greatest effect of session, where participants rapidly become more familiar with their task and are able to improve easily. In later sessions, ceiling performance is being approached, so the difference between subsequent sessions will not be as great and the effect will be reduced.

Results from Session 5

The differences between the variables for the different blocks were compared. The comparisons of greatest interest were between the familiar and "C" test blocks and the performance of the participants in the familiar group on the random blocks compared with the performance of the participants trained solely on random stimuli. The predictions made for the effect of the test are that the participants' performance will be worse on the test blocks than on the familiar blocks presented during the session. This prediction was made as the familiar stimulus which the participants have been trained was to be distorted, to a greater (Random blocks) or lesser ("C" blocks) extent in session 5. It was deemed extremely unlikely that these manipulations would improve performance, hence one-tailed probabilities were used.

Figure 4 shows the data obtained in session 5. The Familiar, "C" and Random entries in the histogram are the averages for each block type collapsed across all the participants in the Familiar group. The control group has been included so that their performance can be compared with that of the Familiar Group on the Random blocks.

When comparing the familiar with the random blocks, all three variables produced significant results. The familiar blocks produced higher scores in the overall percentage correct (T=0, p=0.014) and the percentage of change (T=0, p=0.014) and no change trials correct (T=0, p=0.015). The Random group have been trained on the general task of detecting one square changes within a randomly generated checkerboard. The Familiar group

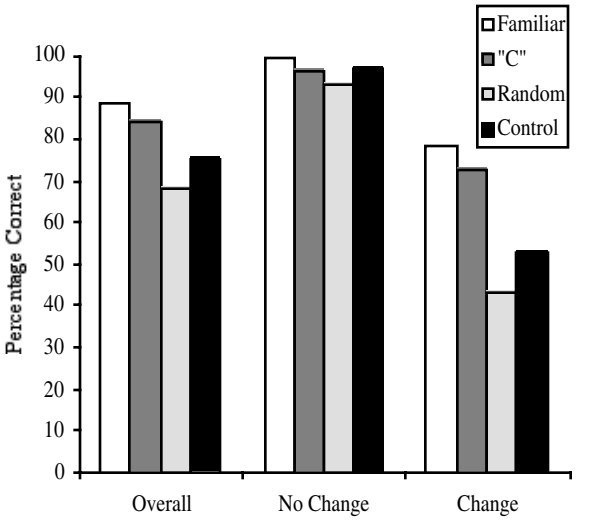


Figure 4: Comparisons between different blocks during Session 5.

have been trained on a task which can be considered as much more specialised. Using the task on which the Random group have been trained allows the comparison between the their performance and the performance attained by the Familiar participants to see if there is any overall advantage of training on a unique stimulus. No significant differences were found for any of the three variables when comparing the performance of the participants in the random group with the scores obtained by the participants in the familiar group on their test blocks with randomly generated stimuli. If anything, the familiar group participants were worse on these trials (Figure 4).

Comparing the results from the familiar and "C" trials, the only significant difference was found between the scores for the percentage of no change trials correct ($T=1$, $p=0.039$) with performance on the familiar trials being better. The other variables were found not to have any significant difference between them. Figure 5 shows the elevated false alarm rate for the "C" condition when compared with the Random and Familiar trials in session 5 and also the control group. The difference between the "C" condition and other blocks is significant in both cases ($T=0$, $p<0.05$). For this analysis, only trials where a response was made are included, thus removing any occasions on which a decision was not made in time.

Figure 5 illustrates the differences in false alarm rates for the four different cases from session 5. These rates for the Familiar participants, using their familiar stimulus, and the control group are on zero, as they registered no false alarms.

Discussion

The effects of session and group have been found to be significant. Although this is not in itself surprising, these two effects had to be established before any further investigations could be carried out. The main effect of group was surprising in its magnitude, with the effect being carried through the sessions. If the participants were allowed to reach asymptotic performance, the prediction from the pilot study is that this significant difference would be maintained. However, this prediction cannot be confirmed as the Random group were not trained for long enough to allow their performance to plateau.

Although the effect of group may have been expected with it being easier to detect changes in a stimulus that is familiar, it could have been the case that there was no difference. If the participants were simply detecting the change between the two checkerboards, then there might have been no difference in the scores. It could have been the case that the unchanging part of the pattern was irrelevant as it was the change between the checkerboards which was being probed. However, the demonstration of a difference between the groups implies that there must be another process at work, apart from the mechanism being used to detect the changes themselves. This position is supported by the finding of no significant difference between the two groups of participants when tested on the random stimuli in session five. Whatever the mechanism

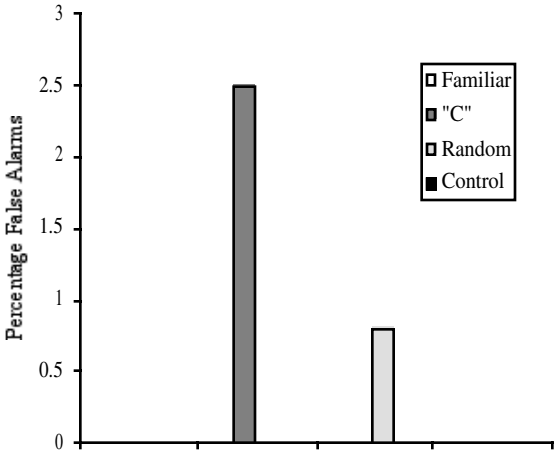


Figure 5: Comparisons of False Alarm rates during Session 5.

favouring the Familiar group, it is specific to the familiar stimuli and not some global strategic advantage developed during training.

The questionnaires administered after the sessions give an insight into how the participants were trying to detect the changes. Every participant reported using a scanning strategy, starting in a particular place for each trial, and then working around the checkerboard, attempting to spot the change. Participants in the random group employed this method throughout their sessions, with no modifications. Participants allocated to the familiar group also used such a strategy, but they reported some modification in later sessions. The scanning became faster, whilst still improving in accuracy between the sessions, there was also a chance to learn about the pattern. They were able to divide the stimulus into sub-patterns and detect changes in these, rather than searching for a single change. This certainly seems to make the detection of differences easier, as performance for the Familiar group was significantly higher for all variables during all five sessions.

The test session produced a significant difference between the percentage of no change trials detected for the "C" blocks and the familiar. In the "C" blocks, there was always a random one-square difference in the checkerboard pairs from the original, familiar stimulus. This additional fixed difference made sure that novelty from the familiar was no longer a necessary signal that there was a difference between the two checkerboards. Many of the participants reported that they had noticed the additional fixed change, and were consciously trying to avoid mistaking it for the actual change between the checkerboards. This result implies that novelty, in addition to a strategic search is central to the task being performed by the participants in the Familiar group. If it were solely a search strategy, then there should be minimal disruption of performance on test blocks which involve a minimal change in the familiar stimulus. The

analysis of the false alarms (Figure 5) supports this position, as there was a greater disruption to the participants on the "C" trials than on the Random trials, where the stimulus was completely different from the one that they were familiar with, and despite the fact that overall performance was significantly lower on the Random trials.

Conclusion

There is good evidence that stimulus familiarity makes it easier to detect a change in that stimulus when compared with participants trained on random stimuli. This is in addition to the necessary attentional requirements reported in Rensink et al. (1997). The ability to detect the change improves with increasing familiarity with the stimulus. However, this ability is limited to that particular stimulus, or ones very close to it, and there is no advantage for other random stimuli of the same type as shown by the test blocks given in session 5.

The mechanism for the detection of the change may be based, at least in part, on novelty. The use of a distracting fixed change induces more false alarms on no change trials than would otherwise be expected. This result indicates that the strategy of scanning the image for change cannot completely account for the enhanced ability to detect the changes. The method for detecting the changes in a given stimulus may involve a combination of better stimulus scanning and the use of novelty to discriminate between familiar stimuli. It is likely that both mechanisms are important in the detection of change. The degree to which each is employed will doubtless depend on both the type and familiarity of the stimulus.

References

Klein, R., Kingstone, A., & Pontefract, A. (1992) Orienting of Visual Attention. In K. Rayner (Ed.) *Eye movements and visual cognition: Scene perception and reading* (pp. 45-65). New York: Springer.

Phillips, W.A. (1974) On the distinction between sensory storage and short-term visual memory. *Perception and Psychophysics, 16,* 283-290

Rensink, R.A., O'Regan, J.K. & Clark J.J. (1997) To See Or Not To See: The Need for Attention to Perceive Changes in Scenes. *Psychological Science, 8,* 368-373

Cue–Readiness in Insight Problem–Solving

Hiroaki Suzuki (susan@ri.aoyama.ac.jp)
Keiga Abe (a1297007@cc.aoyama.ac.jp)
Department of Education, Aoyama Gakuin University
Tokyo 150-8366 JAPAN
Kazuo Hiraki (hiraki@idea.c.u-tokyo.ac.jp)
Department of Systems Science, The University of Tokyo
Tokyo, 153-8902
Michiko Miyazaki (miyazaki@psyche.tp.titech.ac.jp)
Tokyo Institute of Technology
Tokyo, 152-8552 JAPAN

Abstract

This paper explores a mechanism underlying cue-readiness in insight problem-solving. Cue-readiness is concerned with situations where previously neglected information suddenly and unexpectedly becomes illuminative. From the view point of dynamic constraint relaxation theory (Suzuki & Hiraki, 1997), this can be explained by constraint relaxation caused by noticing failures. The theory predicts that constraint violations increase during the problem-solving process, and that a specific combination of constraint violations takes place which leads people to an insight. In this paper, we examined the time-course differences of frequencies of constraint violations, and of sensitivity to the crucial information using a rating task. Although Experiment 1 did not provide supporting evidence, in Experiment 2 we found increased frequency of constraint violations during problem-solving, and that subjects who experienced more failure were more sensitive to crucial information. These results are discussed in terms of other theories of insight.

Insight, one of the most outstanding cognitive activities, is more and more a topic within the scope of rigorous scientific investigation. For the past decade, various approaches have been taken to explore the nature and processes of insight (see, for example, Sternberg & Davidson, 1995).

However, there still remains a mystery. People sometimes find a crucial cue in a relative early stage of problem-solving, but they cannot make use of it. This cue, however, suddenly and unexpectedly becomes illuminative at a certain point, leading problem-solvers to an insight. To put it another way, the same cue has different meanings during the problem-solving process. This can be called "cue-readiness" because it appears analogous to developmental readiness in that the effectiveness of instructional intervention depends on the child's developmental stage.

A good example of the cue-readiness is found in Kaplan and Simon (1990). They used the mutilated checkerboard (MC) puzzle as a material. To solve this puzzle, it is crucial to realize the parity of differently colored squares. In order to control the ease of noticing parity, some subjects in their experiment were given a special board where a word, Bread or Butter, was printed on each square (bread and butter connote parity), instead of colors black or pink. As they predicted, subjects noticed parity more easily and solved the puzzle more quickly. However, they reported one puzzling result. The times from their first mention of parity to the final solution were longer for these subjects than those who were given a standard checkerboard or blank one. While subjects with a Bread–Butter board took 653 s on average to solve the puzzle from their first mention of parity, those with a standard checkerboard took only 110 s.

The problem immediately poses the questions of why people can make use of the crucial cue that they could not do so initially, and what distinguishes the internal states in these two situations.

This problem cannot easily be explained by current theories. Theories based on spread of activation presuppose that the inappropriate problem representations prevent problem-solvers from retrieving an important cue. If this explanation is correct, people could solve the puzzle immediately after noticing the important cue, because the representation of the cue should be activated and the activation spreads over to related information. In the MC puzzle case, subjects could obtain an insight immediately after they mentioned parity.

The idea of the prepared-mind proposed by Seifert et al. (1995) appears to be relevant to the cue-readiness problem. According to them, when people find a standard approach inappropriate, they generate failure indices that mark initial problem solving attempts as unsuccessful. These failure indices are presumed to have the special status in long-term memory, in the sense that they are activated for a longer period than other types of memory traces. In the incubation phase where people stop their initial attempts and are engaged in other activities, a relevant cue is sometimes provided externally, which reminds them of their initial failure and leads them to an AHA experience. We agree that failure and externally provided information play important roles. However, this idea cannot be applied directly to the cue-readiness problem, because their idea deals with a situation where people do not encounter or find crucial information in the initial phase but are given that information externally in the incubation phase. The cue-readiness problem is, however, concerned with a situation where people find crucial information in the initial stage.

In order to deal with the cue-readiness problem, we have developed a dynamic constraint relaxation theory of insight (Suzuki & Hiraki, 1997; Hiraki & Suzuki, 1998).

In the next section, we briefly illustrate the theory.

Dynamic Constraint Relaxation

The dynamic constraint relaxation theory consists of three kinds of constraints (object-level, relational, and goal), and a relaxation mechanism. The main idea is that impasses are formed by these constraints and that qualitative changes are caused probabilistically by the failure-driven incremental relaxation of these constraints.

Constraints

Since it is unlikely that we are equipped with a special cognitive engine for insight problem-solving, it would be desirable that theories of insight do not involve insight-specific mechanisms. One of the most important findings in problem-solving research is that people construct a problem representation consisting of objects, relations, and a goal of the given problem. Reflecting on these findings, we postulate three constraints with objects, relations, and goal. Although the notion of constraints in insight literatures is not new (Isaak & Just, 1995; Knoblich et al., 1999; ; Ohlsson, 1992), our treatment is different from theirs and very similar to analogy (Holyoak & Thagard, 1995).

Object-level constraint There are numerous ways of encoding objects. However, we have a natural tendency to encode them at a basic level (Rosch, 1978). This tendency sometimes becomes an obstacle for insight. For example, in the "Candle" problem, it is well known that people do not notice a pasteboard box of tacks as a holder of the candle. This is because the basic level of a box is "box," not a "solid body" (more abstract) or a "pasteboard box" (more concrete).

We call this tendency the object-level constraint, because it constrains, among possible alternatives, the selection of a specific encoding of a single object. Note here that the constraint is a soft one. It is not that this constraint precludes any other encodings.

Relational constraint Relations define the ways in which objects relate to one another, and each object is assigned a specific role within the relation. Usually, one can relate something to others in various ways. The box in the candle problem, for example, can interact with others in ways of containing, standing on, being thrown to, other objects. However, people usually select the "contain" relation as its default relation.

We call this tendency the relational constraint, because it leads people to select specific relations among numerous alternatives. This constraint is, like object-level constraint, a soft one.

Goal constraint The representation of a goal involves the desired state and evaluation function. This constraint evaluates a match between present and desired states, and gives feedback to the other constraints. Thus, the goal greatly constrains how objects and relations are represented. Although a relation of a candle to other objects is, by default, to light something, a relation such as to glue something by its wax is likely to be selected by the goal constraint.

It is important to note that these constraints interact each other. For example, one reason why the "tacking" relation is selected for the tack is that the basic level encoding of the tack enhances this selection. Another reason is that the goal constraint prevents them from being thrown.

In ordinary problem-solving, these constraints play important roles by eliminating an infinite number of useless representations. However, as noted above, they operate in a harmonious way to form an impasse in insight problem-solving.

Relaxation mechanism

It is important to note that each constraint is not constant during problem-solving, but that its strength changes dynamically. In the course of problem–solving, the mismatch computed by the goal constraint decreases the strengths of initially dominant constraints, which leads to an increase in the probability of constraint-violations. When specific constraint violations occur simultaneously at object-level and relational level, people reach an insight.

In this constraint relaxation process, failure or mismatch detected by the goal constraint plays a key role. A current computational model uses a sort of Q learning algorithm to relax the constraints (Hiraki & Suzuki, 1998). The basic idea is that the strength of the constraint responsible for the failure is reduced to some degree and that the amount of the reduction is distributed to other less dominant constraints by the softmax algorithm (Bridle, 1989).

The dynamic constraint relaxation theory owes much to the multiconstraint theory of analogy (Holyoak & Thagard, 1995). Types of constraints are similar between the two. This is partly because both theories are based on the general characteristics of human problem-solving. However, a crucial difference is that multiconstraint satisfaction often leads to a fruitful analogy, whereas constraint violation leads to an insight in insight problem-solving. Another important difference is that whereas constraint relaxation is purely internal in ARCS and ACME, our theory presumes dynamic interaction with the external environment via feedback.

Previous Studies

We used the T puzzle, similar to the tangram, as material. The goal of this puzzle is to construct the shape of a "T" using four pieces depicted in the left side of Figure 1. At first glance, it appears quite easy to solve, since there are only four pieces and one can easily identify possible positions that some of them should be placed. However, a pilot study, in addition to our own experiences, showed that it is awfully difficult. It usually takes more than half

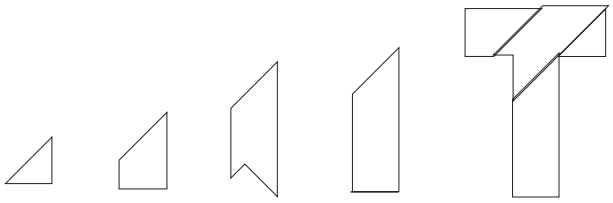


Figure 1: The T puzzle: Construct a shape of "T," using four pieces on the left side.

an hour to solve it spontaneously. Furthermore, more than a few give up trying to solve it.

The difficulties can be explained by the constraints described in the previous section. The object-level constraint in this puzzle is concerned with the preference for how a single piece should be placed, because pieces are objects in the problem representation. People have a strong tendency to place the pentagon piece either horizontally or vertically (Suzuki & Hiraki, 1997). A previous study revealed that subjects placed this piece horizontally or vertically in about 70% of their trials.

The relational constraint in this puzzle is concerned with how one piece is physically connected to one another. The puzzle of this type has an infinite number of relations, because one can produce different patterns by sliding a side of a piece touching another. But, again, people have a strong tendency to connect pieces so as to form a "good" shape with fewer angles. If this constraint actually operates with the goal constraint that evaluates the difference between the current shape and the image of T, it is predicted that people spend most of their time filling the notch of the pentagon. The prediction was confirmed by a previous study which showed more than 70% of the subjects' trials involved notch filling.

Experiment 1

Since our theory predicts that the frequency of constraint violation increases during problem-solving by noticing failure, we analyzed the time-course of constraint violation in Experiment 1. Another dependent variable was subjects' rating score. We used a rating task where subjects evaluated the closeness of various types of combinations of two pieces to the goal. The rating materials were a set of combinations of the pentagon and one of the other pieces, produced by systematically violating the constraints. To control the degrees of relaxation, we divided subjects into two groups, 2-min and 7-min conditions. Subjects were required to solve the puzzle for two or seven minutes, then proceeded to the rating task.

Since subjects in the 7-min condition have failed more often than those in 2-min condition, the theory predicts that the degree of relaxation is higher in the former than in the latter (this is an empirical issue to be examined later). If so, their ratings should be different. According to the theory, the 7-min subjects are more sensitive to crucial information in the rating stimuli than the 2-min subjects. Hence, we expect a statistical interaction between the types of stimuli and before-rating times (2- and 7-min).

Method

Subjects Participants were 33 undergraduate students without any prior experience to solve the T puzzle. They were randomly assigned to 2-min or 7-min condition. We omitted subjects who solved the puzzle before the rating task. Resulting 26 subjects (12 in the 2-min and 14 in the 7-min conditions) were analyzed.

Rating Materials The rating materials consisted of 12 combinations of the pentagon and one of the other pieces (big, small trapezoids, or triangle). These combinations formed four types: O–R– where neither constraints were violated, O–R+ where not object-level, but relational constraint was violated, O+R– where the violating pattern was reversed, and O+R+ where both constraints were relaxed. Since each type had three members depending on which piece was used (big, small trapezoids, or triangle), the total number of rating stimuli was 12.

Procedure The subjects were given the four pieces of the T puzzle and a sheet of paper printed with a 25% reduced-size image of a constructed T. Subjects were asked to construct the shape of "T" using the pieces, with the information about the time allowed to spend before the rating task (2 or 7 minutes).

In the rating task, they were told to rate how close a presented stimulus was to the shape of T with respect to the goal of constructing T, and to click "10" if the stimulus was very close, "0" if it was far from the goal, and other numbers for the intermediary degrees of closeness. Stimulus was presented in a semi-random order that stimuli belonging to the same type were not presented successively. Stimulus presentation time was two seconds, and time for the rating was five seconds.

After completing the rating task, the subjects were asked to resume solving the puzzle. If subjects could not solve the puzzle within 10 minutes from the beginning, the experimenter gave subjects the first hint not to fill the notch of the pentagon (the hint for the violation of the relational constraint). When subjects could not solve the puzzle within five minutes after the first hint, the experimenter gave the second hint not to place the pentagon horizontally or vertically. The entire problem-solving processes were video–taped for the later analysis.

Results and Discussion

To analyze the problem-solving performance, we used a segment as a unit of analysis, in addition to the solution time. A segment is operationally defined as a series of actions that was initiated by physically joining two pieces and terminated by their separation. A segment roughly corresponds to a trial that begins with trying an approach and ends up with noticing failure. It is worth noting that

the notion of segment is not a subjective one, because the definition is based only on physical connections and separations of pieces.

Constraint violation To analyze the time-course of constraint violation, we divided problem-solving processes into four phases, based on the segments (segments after the hints were not included). We counted a segment as a violation of object-level constraint, if the segment did not include the horizontal or vertical placement of the pentagon. We counted a segment as a violation of the relational constraint, if the segment did not have actions to fill the notch of the pentagon by other pieces. Since we found no difference between the two conditions, we merged data obtained from 2- and 7-min conditions. Table 1 shows the proportions of constraint violations in each phase. We conducted one-way ANOVAs for the violation of each constraint separately. We could not find significant time-course difference in the number of segments where the constraints were violated.

Table 1: The percentages of constraint violation in each phase.

	1/4	2/4	3/4	4/4
object-level constraint (%)	24	21	25	25
relational constraint (%)	36	38	38	46

Presenting various types of stimulus did not have a strong effect on the problem-solving performance. The proportions of subjects who solved the puzzle within three minutes after the rating task were 25% in the 2-min condition, and 28.6% in the 7-min condition. These results suggest that majority of the subjects were unable to utilize the useful information presented in the rating task. Additionally, the solution times were not different between the two conditions ($U_{2-min}(12,14) = 66, ns.$).

Rating Before analyzing the rating task data, it is necessary to examine the assumption about constraint-relaxation. Our theory predicts that the more often subjects fail, the more relaxed their constraints are. Hence, we must first examine whether the subjects in the 7-min condition actually failed more often before the rating task than those in the 2-min condition. As we expected, the average number of the segments before the rating task in the 7-min condition was 45.6, while that in the 2-min condition was 17.4 ($t(24) = 7.79, p < 001$).

Table 2 shows the rating score for each type of stimulus. Although the ratings for R–O–, R–O+, R+O– were not different between the two groups, it appears that the 7-min subjects rated the R+O+ type stimuli closer to the goal than the 2-min subjects did. Thus, we conducted a three-way ANOVA to examine the interaction between the types of the stimulus and the conditions. However, the interaction did not reach the significant level ($F(3,72) < 1, ns.$), although there was a main effect of the stimulus types ($F(3,72) = 10.93, p < 005$). Pair-wise comparisons revealed that for both conditions, the R+O+ type was rated closer to the goal than the other types.

Table 2: Mean rating score.

	R–O–	R–O+	R+O–	R+O+
2-min	2.73	2.96	3.90	4.29
7-min	2.83	3.04	3.94	5.35

Experiment 2

The results of Experiment 1 did not support the hypotheses. We found no time-course difference in the frequencies of constraint violations. Furthermore, there was no statistical interaction between the rating scores and the problem-solving time before the rating task. Do these results dismiss the dynamic constraint relaxation theory?

There is, however, the possibility that even for the 7-min subjects, the constraints were less relaxed than expected. According to our theory, one reason is concerned with the goal constraint. As described earlier, the goal constraint plays crucial roles by evaluating the match between the goal and the present state and by giving feedback to the constraints for their relaxation. Actually, previous research revealed that the goal constraint greatly facilitated problem-solving performance (Suzuki et al., 1999). In that experiment, some subjects were given a template sheet printed with an image of a constructed "T," and asked to cover the image by placing the four pieces. Providing the template sheet is expected to facilitate the evaluation of the (mis)match between a current state and the goal. As expected, these subjects solved the puzzle significantly faster than those without the template sheet.

In Experiment 1, subjects were given a sheet of paper printed with an image of "T," but the size was reduced to 25%. In addition, the subjects were not allowed to put the pieces on the sheet. This procedure may cause the goal constraint to operate less effectively. Experiment 2 explores this possibility, by providing the template sheet and instructing subjects to cover the sheet by the pieces.

Method

Subjects Subjects were 20 undergraduate students who had no experience with the "T" puzzle. None of them participated in the previous experiment. These subjects were randomly assigned to either the 1-min or 5-min condition. We omitted three subjects in the 1-min condition and one subject in the 5-min condition who solved the puzzle before the rating task.

Materials The rating materials were 12 combinations of the pentagon and one of the other pieces used in Experiment 1.

Procedure The procedure was basically the same as that of Experiment 1, but there were two modifications.

The first one was to provide subjects with a template sheet printed with an image of "T" and to ask them to cover the image by placing the four pieces. The second one was that the time to solve the puzzle before rating was changed from two and seven to one and five minutes. This was because in a previous study, half of the subjects with the template sheet solved the puzzle within seven minutes.

Results and Discussion

Constraint violation To examine the time course of constraint-violation, we divided the entire problem-solving processes into four phases and counted the number of violations in each quarter, as for Experiment 1. We omitted segments after the hints and merged data obtained from 1- and 5-min conditions. Although the increase of the violation of the relational constraint was not statistically significant ($F(3,48) = 1.07, ns.$), the number of violations of object-level constraints increased dramatically ($F(3,48) = 7.89, p <. 001$). Pair-wise comparisons revealed that the violations of object-level constraints in the final quarter was higher than the others.

The lack of an increase in the number of the relational constraint violations might be due to the fact that the template sheet relaxed the relational constraint from earlier stages. It should be noted that, although the number of constraint violation increased during problem-solving, the constraint violations were observed even in the first quarter. It means that the cue-readiness problem is involved, even when the template sheet was available.

Table 3: The percentages of segments violating the object-level and relational constraints.

	1/4	2/4	3/4	4/4
Object-level constraint (%)	6	19	13	46
Relational constraint (%)	40	41	47	47

Rating As in the previous experiment, we first examined the assumption that 5-min subjects failed more often than 1-min subjects. The average numbers of segments was 50.7 in the 5-min condition and 7.6 in the 1-min condition ($t(11) = 10.15, p <. 001$).

The ratings of each condition were summarized in Table 4. Unlike Experiment 1, we obtained different patterns of ratings. A three-way ANOVA (object-level $\times$ relational constraint $\times$ before-rating time (1- or 5-min)) revealed a significant main effect of the relational constraint($F(1,14) = 41.12, p <. 001$) and interaction between the relational constraint and the time before the rating ($F(1,14) = 10.06, p <. 01$). Although subjects in both conditions gave high rating scores for stimuli that violated the relational constraint, the 5-min subjects gave the highest score for stimuli violating both constraints, whereas the 1-min subjects did so for the stimuli violating only the relational constraint.

Table 4: Mean rating score.

	R–O–	R–O+	R+O–	R+O+
1-min	2.54	2.42	4.38	2.88
5-min	2.58	3	4.04	5.46

General Discussion

In this paper, We propose the dynamic constraint relaxation theory to investigate mechanisms underlying the cue-readiness in insight problem-solving. Our theory assumes that initial impasses are caused by the object-level and relational constraints and that these constraints are gradually relaxed by failures detected by the goal constraint. If our theory is correct, two predictions can be made. First, constraint violations increase during problem-solving processes, because constraints are more relaxed by facing more failures. Second, for the same reason, sudden noticing of crucial information is more likely observed in problem-solvers with more failures than those with fewer. If so, the ratings for constraint-violating stimuli should be different between them.

In order to examine these predictions, we conducted two experiments, using the T puzzle. Subjects' tasks were to solve the puzzle and to rate the closeness of various types of stimulus to the shape of "T." However, we could not obtain any supporting results in Experiment 1. The frequencies of constraint violations did not increase during problem-solving, and the ratings were not statistically different between the 2- and 7-min conditions. However, we found confirming evidence in Experiment 2 where the goal constraint operated more effectively by the template sheet. Violation of the object-level constraint increased when problem-solving proceeded. Furthermore, the ratings of subjects with more failures were different from those with fewer failures in a predicted way.

These results suggest that cue-readiness is caused by constraint relaxation. Due to noticing failure, the probabilities of constraint violations increases during the problem-solving processes, which makes problem-solvers ready to utilize crucial information. Another implication for the problem is that constraint violation at a single level may not be sufficient for insight and it should be coupled with violation at another level.

It is interesting to contrast our theory with a similar view proposed by Knoblich et al. (1999). They have proposed that constraint relaxation and chunk decomposition play key roles in insight problem-solving. Using matchstick arithmetic problems, they found empirical evidence supporting their theory.

Although both theories admit the key roles of constraint relaxation, there are a number of differences between the two. First, constraints used by Knoblich and their colleagues are task-specific. For example, they listed constraints concerning values, operators, and tautology. These constraints are specific to matchstick arithmetic problems, which makes it difficult for their theory

to apply to a large number of insight problems that have no numerical values, mathematical operators, or equal sign.

Second, their theory is not dynamic in the sense that they do not assume any interactions with external environment. In their experiment, subjects were required to *mentally* transform various equations to desired states, which prohibits feedback from the external environment. As Seifert et al. (1995) properly claimed, we obtain information important for modifying our internal states as well as achieving the goal. Therefore, their theory of insight cannot explain findings in the present study, such as the time-course differences of the frequencies of constraint-violation and in the rating patterns observed in Experiment 2.

Third, related to the second, their theory cannot deal with the issue of what relaxes the constraints. They predicted the ease of relaxation based on the notion of the scope of constraints. However, what triggers constraint relaxation remains unanswered. In addition, the scope of the constraint cannot give a principled explanation for the relaxation patterns of the constraints. According to their theory, the relational constraint in our study has wider scope than the object-level one, because the former binds more than one element whereas the latter binds a single element. Thus, their theory predicts that the object-level constraint is more easily relaxed than the relational one. However, we obtained the opposite patterns of relaxation in Experiment 2.

To summarize, we agree that constraints forms an impasse and that insight is achieved by constraint relaxation, but oppose their notion of purely "cognitive" insight as well as task-specificity of constraints.

Acknowledgment

This research was supported in part by Grant-in-Aid for Scientific Research (C)(No. 10610082 and 12680390). We thank Steven Phillips for his helpful proofreading and comments, and Yasuhiro Hiraoka for his help in conducting the experiments.

References

Bridle, J. (1989) Probabilistic interpretation of feedforward classification network outputs with relationships to statistical pattern recognition. In F. Fogelman-Soulie & J. Herault (Eds.) *Neuro-computing: Algorithms, Architectures*. Springer-Verlag.

Hiraki, K. & Suzuki, H. (1998) Dynamic constraint relaxation as a theory of insight. *Bulletin of Cognitive Science*, **5**, 69 – 79. (In Japanese with an English abstract)

Holyoak, K. J. & Thagard, P. (1995) *Mental Leaps: Analogy in Creative Thought*. Cambridge, MA: MIT.

Isaak, M. I. and Just, M. A. (1995) Constraints on thinking in insight and invention. In R. J. Sternberg and J. E. Davidson (Eds.) *The Nature of Insight*. Cambridge, MA: MIT.

Kaplan, C. A. and Simon, H. A. (1990) In search of insight. *Cognitive Psychology*, **22**, 374 – 419.

Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999) Constraint relaxation and chunk decomposition in insight problem-solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **25**, 1534 – 1555.

Ohlsson, S. (1992) Information processing explanations of insight and related phenomena. In M. T. Keane and K. J. Gilhooly (Eds.) *Advances in the Psychology of Thinking, vol. 1*, Hertfordshire, UK: Harvester.

Rosch, E. (1978) Principles of categorization. In E. Rosch & B. Lloyd (Eds.) *Cognition and Categorization*. Hillsdale, NJ: Erlbaum.

Seifert, C. M., Meyer, D. E., Davidson, N., Patalano, A. L. & Yaniv, I. (1995) Demystification of cognitive insight: Opportunistic assimilation and the prepared-mind perspective. In R. J. Sternberg & J. E. Davidson (Eds.) *The Nature of Insight*. Cambridge, MA: MIT.

Suzuki, H. & Hiraki, K. (1997) *Constraints and their relaxation in the processes of insight*. ETL Technical Report, TR–97–13.

Suzuki, H., Miyazaki, M. & Hiraki, K. (1999) Goal constraints in insight problem-solving. Proceedings of the Second International Conference on Cognitive Science, 159 – 164.

Sternberg, R. J. and Davidson, J. E. (1995) *The Nature of Insight*. Cambridge, MA: MIT.

Extending the Past-tense Debate: a Model of the German Plural

Niels A. Taatgen (niels@ai.rug.nl)
Artificial Intelligence, University of Groningen
Grote Kruisstraat 2/1, 9712 TS Groningen, the Netherlands

Abstract

One of the phenomena that has been studied extensively in cognitive science is learning the English past tense. Many models have been made of the characteristic U-shape in performance on irregular verbs during development. An important test case for such models is whether they can be extended to other examples of inflection. A case that is often quoted as particularly tough is the German plural. In the present study, an ACT-R model of the past tense is applied to the German plural. The model not only successfully learns the default rule, but also exhibits some other characteristics of the German plural.

Introduction

Learning the English past tense has been one of the central topics of debate in cognitive science since McClelland and Rumelhart published their original neural network model in 1986. The phenomenon is very simple. English verbs can be broken down into two categories: regular and irregular verbs. The past tense of a regular verb can be obtained by simply adding *-ed* to the stem. Irregular verbs on the other hand are unsystematic: each verb has a unique inflection. When children have to learn the inflection of the past tense, they go through three stages. In the first stage their use of the past tense is infrequent, but when they use the past tense they do so correctly. In the second stage they use the past tense more often, but they start overregularizing the irregular verbs. So instead of saying *broke*, they now say **breaked*. On the other hand, inflection of regular verbs increases dramatically, indicating that the child has somehow learned the general regular pattern. In the third stage, they inflect irregular verbs correctly again. This pattern of learning is often referred to as U-shaped learning.

Although learning the past tense seems to be a rather simple problem, it nevertheless encompasses a number of issues in language acquisition and learning in general. Apparently the past tense has two aspects: on the one hand there is a general rule, and on the other hand there is set of exceptions. Children are able to learn both aspects, and the phenomenon of U-shaped learning seems to implicate that children learn the general rule in stage 2. The important point McClelland and Rumelhart make is that this does not necessarily imply that this knowledge is actually represented as a rule in the cognitive system: their neural network model has no separate store for rules, but it nevertheless exhibits rule-like behavior in the form of U-shaped learning. Ever since their original model, the neural network approach has been challenged (e.g., Pinker & Prince, 1988), improved (e.g., Plunkett & Marchman, 1991), challenged again (e.g., Marcus, 1995) and improved again (e.g., Plunkett & Juola, 1999). I would like to highlight two unresolved issues in this debate, because they will be addressed here. The first issue is feedback. A well-known fact in language acquisition is that children do not rely on feedback on their own production of language (at least with respect to syntax), simply because they do not receive any (Pinker, 1984). Although this problem is addressed by some modelers (e.g., Plunkett & Juola, 1999), its resolution is not entirely satisfactory: the assumption is that learning takes place while children perceive past tenses, and not while they actually produce past tenses. This idea is at odds with the picture of skill acquisition in general, where practice is considered as a main means of learning. A second issue is the frequency of the regular cases. In English, most verbs are regular. This fact is essential for neural network models, as they need to be presented with regular cases at least 50% of the time (Marcus, 1995). This is already slightly problematic in English, as the token-frequency of regular verbs, how often a verb is actually used in language, is only around 30% (irregular verbs are just used much more often than regulars). Connectionist modelers have therefore introduced the input/uptake distinction: not every word that is perceived is presented to the network. This assumption becomes especially problematic if regular forms are much more rare. An example of inflection where the regular form is very rare is the German plural.

The German Plural

German has five different suffixes to mark plurality of a noun: zero (no suffix), *-(e)n*, *-e*, *-er* and *-s*. Moreover, the stem-vowel sometimes receives an Umlaut (¨), something we will ignore for the present. The plural is almost always indicated by suffixation: there are only a few exceptions, mainly words derived from Latin (e.g., *Thema-Themen*). Careful analysis of these suffixes has revealed that the *-s* suffix is actually the default rule (Marcus, Brinkmann, Clahsen, Wiese & Pinker, 1995). Interestingly enough, this suffix is also the least frequent of all five, both in *type-frequency* (how many words are there) and *token-frequency* (how often are they used). Marcus et al. estimate the type frequency of nouns ending in *-s* at 4%, and the token frequency at only 2%. It appears however that at least some of the other suf-

fixes are somehow tied up by additional constraints: for example, zero and *-er* are never used for feminine words, *-e* cannot be used if the stem already ends with and *e*, etc.

The combination of a default rule that is based on very low frequencies and the fact that there is no feedback on production makes it very hard to understand how the default rule can be learned at all. If there is no feedback, the cognitive system has to construct its own language based on perceived inputs from the environment. But why would the cognitive system always elect to use *-s* as a default rule, while there are other options as well?

A possible source of information on this topic is to look at children, and examine the type of errors they make. Marcus et al. (1995) quote a number of studies that indicate children overregularize using the *-s* suffix (in 10-15% of the opportunities), although it is not the most common overregularization (*-(e)n* is the most common overregularization). This pattern is similar to the English past tense, which has been studied much more extensively, except that in English the default rule is also the dominant source of overregularization.

Regular versus irregular inflection is often characterized by competition between a rule and exceptions. It appears that in the case of the German plural there is also competition among rules, a competition the *-s* rule eventually wins.

To summarize, the German plural is in some sense similar to the English past tense, but more complicated. There is competition among candidate rules, while the English past tense has only one apparent candidate rule, and the eventual rule is based on nouns that have a low frequency, as opposed to the high frequency of regular English verbs. The German plural is therefore an interesting test case for existing models of the past tense in English: can these models successfully account for the German plural as well? Taatgen and Anderson (submitted) developed a model for the English past tense, the present study will show it is extendable to the German plural as well, without many modifications.

Towards a General Model of Regular and Irregular Inflection

The Taatgen and Anderson (submitted) model of learning the past tense is based on the ACT-R architecture (Anderson & Lebiere, 1998). It is a so-called dual-representation model, so it separately represents examples and rules, corresponding to ACT-R's declarative chunks and procedural rules. Although two representations make the model weaker than neural network models that use only one type of representation, it does not need a number of assumptions neural network models need. The ACT-R model does not need a specific input regimen, in which the vocabulary is gradually increased. Neural network models typically start training with a small set of words, in the order of 10 to 20. This number is increased during training. A problem with this approach is that the way in which the input-set is increased can be manipulated to get the desired outcome. The ACT-R model on the other hand can be trained on the full vocabulary right from the start, and with the exact token frequencies as in normal language.

A second advantage of the ACT-R model is that it can learn without feedback on its own performance. The model assumes examples of past tenses are perceived and stored in declarative memory, and that no feedback is given on production. The only feedback the model uses when it produces language is its internal feedback: the effort production took. It will prefer strategies that take the least effort, as opposed to strategies that produce the right answer (as it has no way of knowing what the right answer is). Neural networks have to make additional assumptions to account for the lack of feedback. As a neural net needs to know the correct answer to adjust its weights, it has to learn during the perception of language instead of during production. It is assumed that once an inflected form is analyzed, it is "recreated" by a hypothesis generator. To quote Plunkett and Juola (1999):

> The child is continually taking in word tokens and comparing the words actually heard (e.g., "went") to the tokens that the child's hypothesis generator would have expected to produce as inflected forms of a given stem; when they differ, this provides evidence to the child that the hypotheses are wrong and should be modified. (p. 466)

Although this view on language acquisition is not necessarily false, there is no clear evidence for it, and must be considered as an extra assumption. To summarize, although neural network models need only one form of representation, and are stronger theories in that respect, they are weaker with respect to the shape of the input and the organization of feedback.

One of the main claims of cognitive modeling is that models can be generalized to other tasks and contexts. A model of the past tense should be a stepping stone towards a more general model of regular and irregular inflection in different languages. The first step is discussed in this paper: the German plural. Before I will discuss the model, I will briefly explain a few relevant ACT-R aspects.

Rules and Examples in ACT-R

According to the ACT-R theory and architecture (Anderson & Lebiere, 1998) human memory consists of two long-term stores: a declarative memory and a procedural memory. As ACT-R is a hybrid architecture, representations in both memory systems have symbolic and subsymbolic aspects.

Declarative memory is used to store facts, goals and perceptual information. For the purposes of the model, it will store the words in the vocabulary, and examples of how an inflected form, in this case the plural form, is constructed. The declarative memory may store the fact (called a *chunk*) that "Jahr" (year) is a noun, and that "Jahre" is the plural of "Jahr". Declarative memory may contain false facts along-

side true facts, so it may have a chunk "Jahr"-"Jahren" as well as the correct chunk. Each chunk has an activation value, that represents the log odds that the chunk will be needed in the current context. So ACT-R doesn't really care about truth although true facts are probably more often needed than false facts.

For the purpose of the current model, the main determiner of the activation value is repetition. If a certain chunk is retrieved often from memory, or is perceived often, its activation value will be high. The main effect of activation for the present model is that chunks whose activation is too low cannot be retrieved from memory. Also, if two or more chunks are candidates for retrieval at the same time, the chunk with the highest activation is chosen (more precisely: has the highest probability of being chosen as noise is added to the process).

Chunks cannot act by themselves, they need production rules from procedural memory for their application. In order to use a chunk, a production rule has to be invoked that retrieves it from declarative memory and does something with it. Since ACT-R is a goal-driven theory, chunks are always retrieved to achieve some sort of goal. In the context of inflection of words the goal is simple: given the stem of a word, produce the proper inflection. One strategy to produce a certain inflection is to just retrieve it from declarative memory, using a production rule like:

```
IF    the goal is to produce a certain
      inflection of a word
      AND there is a chunk that specifies this
      inflection for that word
THEN  the set the answer of the goal
      to that inflected form
```

If the goal is to produce the plural of a certain word, this production rule will attempt to retrieve a chunk from declarative memory that specifies what the plural is of that word. Of course this production rule will only be successful if such a chunk is present and its activation is high enough.

The behavior of production rules is also governed by subsymbolic parameters. Each production rule has some real-parameters associated with it that estimate its expected outcome. This expected outcome is calculated from estimates of the cost (in time) and probability of reaching the goal if that production rule is chosen. The unit of cost in ACT-R is time. ACT-R's learning mechanisms constantly update these estimates based on experience. If multiple production rules are applicable for a certain goal, the production rule is selected with the highest expected outcome. This is again a noisy process, so the rule with the highest expected gain only has the highest probability of being selected first. If a rule is selected and subsequently fails, the next best rule is tried. For example, if the example rule above fails to retrieve a chunk that specifies the inflected form, the next best rule will be tried.

New rules are learned through a process of specialization and compilation. In the present model, this process will specialize the general strategy of analogy into specific rules for inflection[1]. I will discuss the details of this process later in the paper.

The Model of the German Plural

Prior knowledge in the model

The model starts out with a set of general problem-solving strategies that are often used in ACT-R models:

1. **Retrieval**. A general strategy for problem solving is to search declarative memory for a case that is identical to the case at hand. If an identical case can be found, it immediately produces the answer, else one of the other strategies is tried.
2. **Analogy**. Another general strategy is to look for a case that is similar to the current problem. After a suitable example has been found, a mapping has to be found between the problem and the answer in the example. This mapping is then applied to the case at hand. In the present model, an arbitrary case of a plural noun is retrieved from declarative memory, after which very simple pattern-matching productions identify the mapping and apply it to the current noun.
3. **Do nothing**. A third strategy is to just do nothing. This strategy is not as stupid as it sounds, as some problems may not be worthwhile solving as long as no additional knowledge is available. The result of doing nothing is that the singular form will be used instead of the plural.

Input

The CELEX database was used to select the 538 most frequent German nouns. After removing the nouns that have no plural, 472 nouns were left. The frequencies of the plural forms are in Table 1. Note that the frequencies for the *-s* suffix are even lower than the Marcus et al. estimates, probably because only high frequency words were selected.

Table 1: Frequencies of suffixation in the CELEX sample

Suffix	Type frequency	Token frequency
-(e)n	48%	50%
-e	34%	35%
zero	11%	8%
-er	5%	7%
-s	1.3%	1%
other	1%	0.4%

To simulate a child's perception and production of plurals, the following input regimen for the model is used: every 2000 seconds, one word is randomly drawn from the set of

1. The process of proceduralization used in this model is not part of ACT-R 4.0, the current version of ACT-R, but is part of a proposal for the next version of the architecture

472 words, based on the token frequency of the word. The model has to produce the plural of this word, simulation production by child. As token frequencies are used, high-frequency words are drawn more often, in the same proportion as they occur in the corpus. Also, every 2000 seconds two random plurals, drawn in the same manner, are added to declarative memory, reflecting perception from the outside world.

Learning

During the simulation, several learning processes influence the behavior of the model, from symbolic to subsymbolic and from declarative to procedural. Table 2 summarizes the different learning mechanisms.

Table 2: Learning mechanisms and their effects

Type of learning	**Effect on the model**
Declarative Symbolic	Examples of plurals are added to memory. Examples are perceived in the environment, and produced by the model itself.
Declarative Subsymbolic	Examples that occur often or are retrieved often are more readily available in memory, as they receive a higher activation. Low-activation examples cannot be retrieved.
Procedural Symbolic	Rules are learned to add specific suffixes to the stem in order to create a plural.
Procedural Subsymbolic	The expected gain of each strategy is estimated based on experience: the rule that takes the least effort to produce a plural is favored.

A first aspect of learning is that new production rules are learned that add the different suffixes to the stems. These rules are learned by specializing analogy. Analogy can be characterized by two steps: retrieving an example from declarative memory and applying this example to the new case. Proceduralization eliminates the retrieval of the example, and substitutes variables in the rules with a certain example. It then combines the two steps in a single step. The result is a rule that approximately acts like analogy, but always with the same example. In the case of inflection, the suffix of the retrieved example determines what the new rule will do: if the example has an *-e* suffix, the new rule will always add the *-e* suffix to produce a plural.

A second procedural aspect of learning is that rules compete. The three strategies mentioned before, together with the rules that proceduralization produces, all compete in producing an inflection. Although they do not receive feedback whether what they produce is correct, they do receive feedback on how much effort it took to produce an inflection. This effort can be influenced by many factors. For example, if the retrieval rule fails to find an example, another strategy has to be tried afterwards, increasing the average effort of retrieval. If a rule produces a suffix that is long to pronounce, using that strategy takes more effort than a short suffix. With respect to this pronunciation effort, the model assumes that using a suffix implies some extra effort. It shares this assumption with the past tense model. Moreover, it assumes the *-s* suffix takes slightly less effort than the other suffixes, because *-s* suffix is just an additional phoneme, while the other suffixes are extra syllables. This is an important assumption, because it will be the main reason why the *s*-suffixation rule will eventually dominate other suffixation rules.

Learning in declarative memory also plays a key role in this model. On the symbolic level, examples of past tenses are constantly added to memory, by perceiving them in the outside world, but also by producing them. The fact that an example is in declarative memory does not guarantee that it can be retrieved. This is where the subsymbolic level is important: activation decays with time, making the example irretrievable. Another aspect of activation is to decide in the case of multiple choices: if two chunks match, the chunk with the highest activation is chosen.

It is important to note than whether or not a produced plural is correct has no impact on the learning. Correctness only plays a role in the examples that the model perceives in the world, but even there an occasional error will not disrupt performance.

Results of the Model

The model was run for 80000 trials, or slightly over 60 simulated months. Figure 1 shows the expected gains of the different rules. Remember that the rule with the highest expected gain is generally tried first, and if it fails the next best rule is tried (although noise may change the order from time to time). Right from the start, retrieval is the dominant strategy. Its expected gain improves quickly as more and more examples are learned. Retrieval is not always successful, so the order of the remainder of the rules is especially important. The rules for the zero, *-e*, *-(e)n* and *-er* suffixes are learned very early in the simulation, and appear to be reasonably productive, as they pass both the do-nothing and the analogy strategy around month 5. Only after 10 months in the simulation the *-s* rule is learned, due to the fact that its occurrences and therefore the opportunities for generalization are rare. Once the *-s* rule is learned, however, it quickly dominates the earlier suffixation rules due to its pronunciation advantage.

The expected gains of the rules have a direct impact on the performance of the model, depicted in Figure 2. As the expected gain of the retrieval-rule increases, the proportion of correct responses also increases (Figure 2a). When after

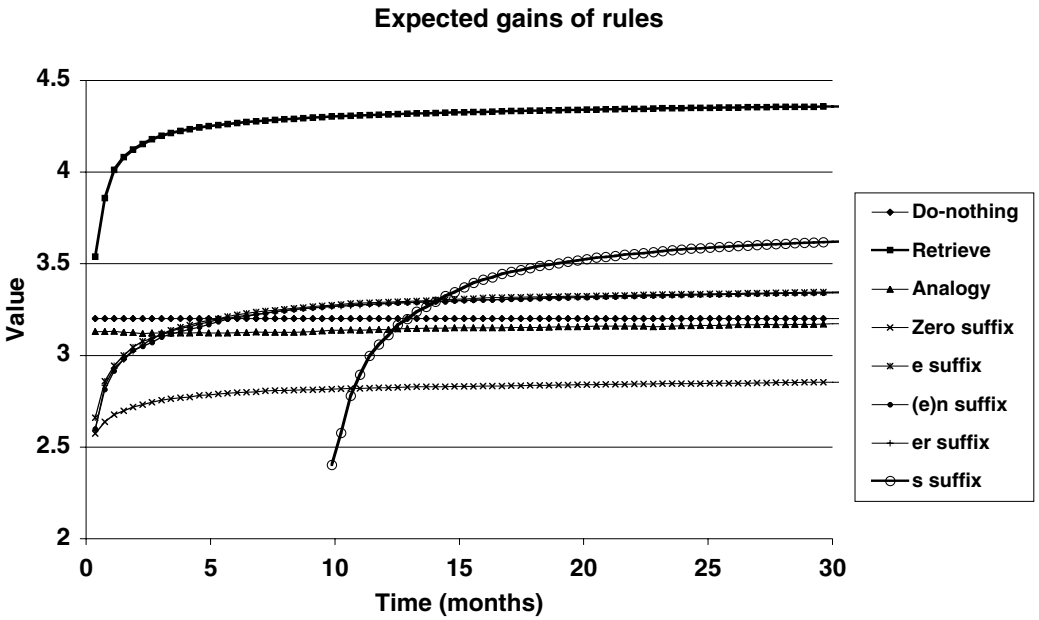


Figure 1: Expected Gains of the different strategies during the first 30 months of the simulation

10 months the *-s* rule is learned, the model starts making errors by adding it to other stems.

Figure 2b shows all the errors that the model makes: at the start of the simulation, errors are dominated by the "do-nothing" rule (producing just stem, so a zero-suffix), as very few plurals are yet known. Soon afterwards, the rules for *-(e)n*, *-e* and *-er* are learned, so they dominate the errors. After month 10, the rule for *-s* dominates the errors, although other errors are still made due to noise, and retrieval of past errors. If we extrapolate the results presented here towards adulthood, the dominant strategy will be to retrieve the plural form from memory. If that process fails, the rules that add the *-s* suffix will be used: exactly what one would expect from a default rule.

Discussion

The present model shows that the original Taatgen and Anderson (submitted) model of the English past tense can be extended to the German plural without modifications. The model is able to learn the default rule despite the low incidence of examples, which is an important problem for other models. But how well does the model fit the data? Unfortunately, the data on the German plural is not as extensive as data on the English past tense.

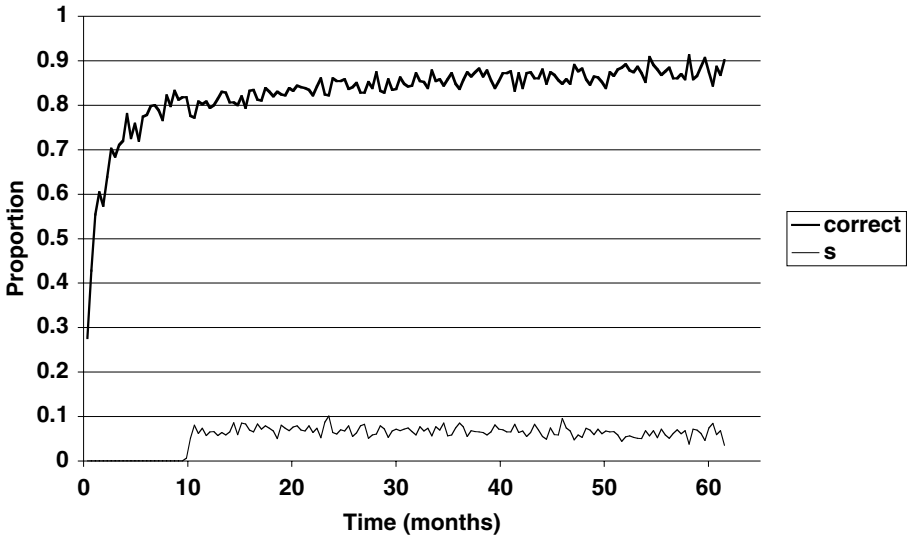
The basic facts from Marcus et al. (1995) are that German children overregularize by using the *-s* suffix, but also by using other suffixes like *-(e)n*. Also, German children make much more errors in the plural (Marcus et al. quote percentages between 11% and 25% for just the *-s* overregularization) than English children with the past tense (usually less than 10%). Both these facts are supported by the model. The fact that retrieval is the dominant strategy implies that rules play only a minor role in inflection, and that this role is most important during the learning phase, when not all inflected forms are memorized yet. In English, the regular rule for the past tense leads to reasonable performance, as the majority of the verbs is regular. In German, however, the regular rules will generally not produce correct behavior, so it is no surprise children make many errors.

The present model operates on a rather global level, largely ignoring some issues concerning phonetics and gender. Obviously, the *-e* suffix cannot be used when the stem already ends with an *-e*. This would make the *-e* rule less attractive, as it will sometimes fail. Gender may also place additional constraints on certain rules. Furthermore, due to phonological constraints, it is sometimes necessary to add the Umlaut to the vowel in the stem with certain suffixes. The *-s* suffix is free of all these constraints, and can in principle be applied in all cases. Together with the fact that it is also the shortest suffix, this makes the rule the most attractive one, despite its low incidence.

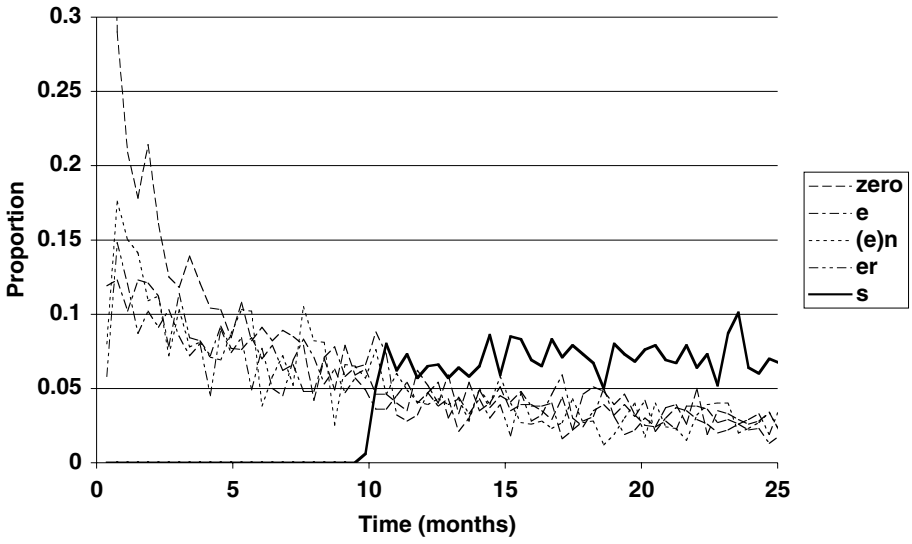
Despite the fact that it largely ignores some of the low-level details, this model demonstrates how generalization in language acquisition can be explained in the absence of feedback. Although it does not solve the learnability problem in language, it nevertheless points at a different source of feedback that may play a role in different areas of language acquisition as well: internal feedback that is not based on the correctness of the produced utterance, but based on the amount of effort it took to produce the utterance.

Acknowledgments

I would like to thank Jack Hoeksema for help in consulting the CELEX database.



(a)



(b)

Figure 2: Performance of the model (a) proportion of correct responses, and overregularization errors with the *-s* suffix (b) proportions of all errors in the first 25 months of the simulation.

References

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Marcus, G. F. (1995). The acquisition of the English past tense in children and multilayered connectionist networks. *Cognition*, *56*, 271-279.

Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., & Pinker, S. (1995). German inflection: the exception that proves the rule. *Cognitive Psychology*, *29*, 189-256.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a distributed processing model of language acquisition. *Cognition*, *28*, 73-193.

Plunkett, K. & Juola, P. (1999). A connectionist model of English past tense and plural morphology. *Cognitive Science*, *23*, 463-490.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition*, *38*, 43-102.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tense of English verbs. In J.L. McClelland & D.E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 216-271). Cambridge, MA: MIT Press.

Taatgen, N.A. & Anderson, J.R. (submitted). Why do children learn to say "Broke"? A model of learning the past tense without feedback. Prepublication available on-line: http://ai.rug.nl/prepublications/prepubsTCW-2000-9.pdf

The modality effect in multimedia instructions

Huib K. Tabbers (huib.tabbers@ou.nl)
Rob L. Martens (rob.martens@ou.nl)
Jeroen J. G. van Merriënboer (jeroen.vanmerrienboer@ou.nl)
Open University of the Netherlands: Educational Technology Expertise Centre; P.O. Box 2960
NL-6401 DL Heerlen, The Netherlands

Abstract

The influence of presentation format on the effectiveness of multimedia instructions was investigated. According to Cognitive Load Theory (Sweller, Van Merriënboer & Paas, 1998) and Mayer's theory of multimedia learning (Moreno & Mayer, 1999), replacing visual text with audio will decrease working memory load and improve learning (modality effect). This hypothesis was tested in two experiments in which students studied multimedia instructions on an instructional design model. The students reported the mental effort spent on the instructions, and made a retention and a transfer test after the instructions. The results show that replacing text with audio is only effective when multimedia instructions are system-paced.

Introduction

Guidelines for the design of multimedia instructions are often based on intuition and practical experience rather than on the results of experimental research (Park & Hannafin, 1994). However, two recent lines of research that have yielded some interesting results are the work by John Sweller and his colleagues on Cognitive Load Theory (Sweller, 1988; Sweller, van Merriënboer & Paas, 1998), and the experiments carried out by Richard Mayer and his colleagues on multimedia learning (for an overview, see Moreno & Mayer, 1999). Both researchers claim that multimedia instructions consisting of verbal and pictorial information, like for example a picture of a machine and a text about its functioning, place a high demand on working memory resources, because the learner has to switch between text and picture in order to integrate them mentally. An interesting finding in their research is that this memory load can be reduced by presenting the verbal information auditorily instead of visually. They call this phenomenon the *modality effect* or *modality principle*. The explanation they give is based on the working memory model of Baddeley (1992). In his model, working memory has two modality-specific slave systems: one for processing visual and spatial information and one for acoustic information. When information is presented in two sensory modalities (visual and auditory) rather than one, both slave systems are addressed and total working memory capacity is increased. So relative to the available resources, the memory load of the multimedia instructions is reduced, leaving more space for the actual learning process.

Sweller and Mayer have demonstrated the superiority of audio over written or on-screen text in a number of experiments. For example, Jeung, Chandler and Sweller (1997) and Mousavi, Low and Sweller (1995) showed that students receiving multimedia instructions with audio spent less time on subsequent problem solving compared to students receiving visual-only instructions. Furthermore, students in experiments by Kalyuga, Chandler and Sweller (1999) and Tindall-Ford, Chandler and Sweller (1997) reported less mental effort during instruction and attained higher test scores, while in the studies by Mayer and Moreno (1998; 1999) students had higher scores on retention, transfer and matching tests. In one experiment, Moreno and Mayer (1999) even used instructions in which the animation and the accompanying text were presented sequentially instead of simultaneously. Despite the temporal detachment of text and picture, bimodal instructions still proved to be superior to visual-only instructions. This shows that the modality effect seems to be at least for some part the result of an increase in available memory resources.

Based on these results, Sweller and Mayer strongly advocate the use of audio in multimedia instructions. However, one limitation of the above-mentioned studies is that they all deal with short multimedia instructions on well-defined technical subjects like geometry (Mousavi et al., 1995; Jeung et al., 1997), scientific explanations of how lightning develops (Mayer & Moreno, 1998; Moreno & Mayer, 1999) and electrical engineering (Kalyuga et al., 1999; Tindall-Ford et al., 1997). This raises the question how powerful the modality effect actually is. Can it also be demonstrated with multimedia instructions that are outside the technical domain and are of greater length? This question is dealt with in the first experiment of this study.

The second issue that can be raised given the evidence so far, is that the results can be explained in more than one way. For example, Jeung et al. (1997), Mousavi et al. (1995) and Tindall-Ford et al. (1997) used visual-only instructions in which the complete explanatory text was printed next to the diagram and

compared it to instructions in which the students only saw the picture and could listen to the explanation. That means that they not only replaced visual text with audio, but also reduced the visual search necessary to link the right parts of the text with the right parts of the diagram. So in their experiments, the difference in effectiveness between bimodal and visual-only instructions could also be attributed to the difference in visual complexity.

On the other hand, Mayer and Moreno (1998; 1999) and Kalyuga et al. (1999) cut their explanatory texts in smaller pieces and still found a modality effect. However, in their experiments the instructions were presented as system-paced animations. The time a student could study a picture and its accompanying texts was determined by the speed of the narration in the bimodal condition. The learners in the bimodal condition could use this limited period of time effectively because they could look at the picture and listen to the text at the same time. The learners in the visual-only condition on the other hand had to spend part of their time in a process of skipping back and forth between text and picture in order to integrate them mentally. We question if the modality effect will still appear if you give the students in the visual condition more time to relate the text to the picture. This issue is dealt with in the second experiment.

Experiment 1

The aim of our first experiment was to see if we could replicate the modality effect using longer multimedia instructions in a non-technical subject domain. For this purpose we developed web-based instructions on an instructional design model. The material mainly consisted of diagrams with explanatory texts. Jeung et al. (1997) showed that replacing visual text with audio does not always improve the effectiveness of multimedia instructions, especially when using pictures with a high visual complexity. They argued that the visual search needed to find the part of the picture the text is referring to increases the memory load. After adding visual cues to the pictures in the form of electronic flashing they regained the modality effect. In our experiment we used colour coding as a means of preventing unnecessary visual search.

The hypothesis that follows from Cognitive Load Theory and Mayer's work on multimedia learning is that presenting the texts accompanying the diagrams as audio will decrease the working memory load of the instructions. Therefore we divided the students in two groups, one receiving bimodal instructions (the audio group) and one receiving visual-only instructions (the visual group), and measured the mental effort spent on the instructions. Paas and Van Merriënboer (1994) argue that mental effort is just one dimension of cognitive load that is not only influenced by task-characteristics but also by subject characteristics like prior knowledge and subject-task interactions like motivation. We tried to exclude any of these effects by randomization of our subjects over the groups, so that differences in mental effort scores could be attributed to the differences in presentation format.

It is also possible that the freed working memory resources are used for the learning process itself and no differences in mental effort are reported. Therefore we also looked at the extent in which students could recall elements of the design model in a retention test, and at the extent in which they could apply the model in a new situation with a transfer test, to see if there was any difference in performance. Finally, we also measured the mental effort spent on both tests.

Method

Participants The participants were 41 students from a Teacher Training College for Primary Education in Heerlen, the Netherlands (20 second-years and 21 third-years; age between 18 and 24; 11 males and 30 females). They had applied on a voluntary base and were paid forty guilders for their participation. During their studies, the students hadn't had any lessons on instructional design models. Twenty participants were randomly assigned to the visual group and 21 to the audio group.

Materials We developed web-based multimedia instructions on the Four Component Instructional Design (4C/ID)-model of Van Merriënboer (1997). This model describes a design strategy for the training of complex cognitive skills. The instructions focused on the question how to develop a blueprint for a training program based on the skills-hierarchy of a complex skill. The instructional website started with a short textual introduction to the model. Subsequently, the design strategy of the 4C/ID-model was demonstrated in a series of eleven diagrams representing skill hierarchies and sequences of learning tasks. These diagrams formed two worked-out examples and a general explanation of the strategy. The first example consisted of six diagrams that showed the different stages in developing a blueprint for the training of the complex skill *doing experimental research* (see Figure 1a and 1b for screen examples). The second worked-out example consisted of three diagrams showing the same process for the complex skill *designing a house*, and finally the general strategy of the 4C/ID-model was explained in the last two diagrams.

All eleven diagrams were accompanied by a textual explanation on how the model was applied in the specific situation. These explanatory texts were cut up into smaller pieces of only one or two sentences long, in such a way that each piece of text referred to a specific part of the diagram. Moreover, these parts were

coloured bright red in the diagram to prevent any unnecessary visual search. So while studying a diagram, only the accompanying text and the colour coding changed, not the diagram itself.

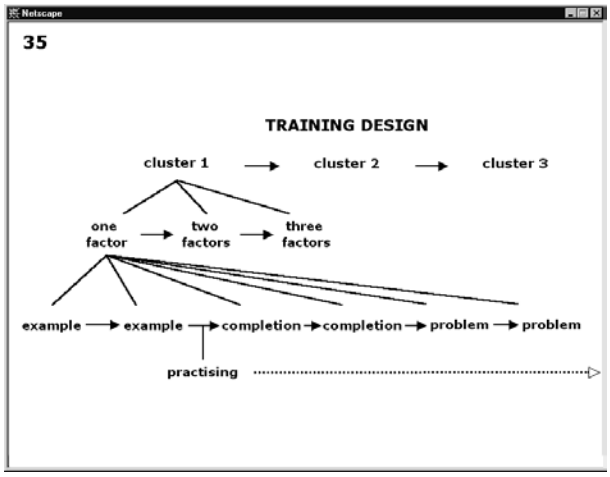


Figure 1a: Screen example of the audio version of the multimedia instructions.

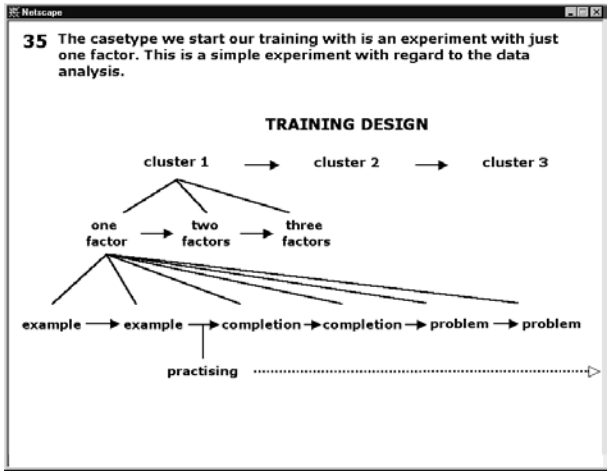


Figure 1b: Screen example of the visual text version of the multimedia instructions.

Two versions of the instructional website were created that differed in the way the texts accompanying the diagrams were presented. In the audio version (Figure 1a), students could listen to the pieces of explanatory text that accompanied a diagram through a headphone. Three seconds after the audio had finished playing, the colour coding in the diagram changed and the next piece of audio started. In the visual text version (Figure 1b) the pieces of explanatory text were depicted right above the diagrams. After the same period of time as in the audio condition, the colour coding in the diagram changed and a new piece of text appeared above the diagram. The time it took to study all eleven diagrams was about 30 minutes

After each diagram, a separate page followed with a nine-point rating scale on which the students could rate the mental effort they had spent on the instructions. This scale was developed by Paas and others (Paas & van Merriënboer, 1994; Paas, van Merriënboer & Adam, 1994). When a student clicked on one of the nine options, the program automatically continued with the next diagram. The average score on the eleven rating scales was taken as a measure of mental effort during instructions.

The retention test consisted of two paper-and-pencil tests, one of 30 and one of 20 multiple-choice items. The 30-item test contained only verbal statements, while the 20-item test combined verbal statements with small parts of diagrams. All items were statements about the 4C/ID-model like “A macro-sequence in the 4C/ID-model is a series of subskills in a cluster”, or “According to the 4C/ID-model, the same subskills can be trained in more than one learning task”, and the students could choose between *correct*, *incorrect* or *I don't know*. Each right answer yielded one point. The retention score was calculated by taking the sum of the scores on all fifty items (Cronbach's alpha = .74). A nine-point rating scale similar to the ones used in the instructions followed both multiple-choice tests. The average score on both scales was taken as a measure of the mental effort spent on the retention test.

The transfer test was also a paper-and pencil test that contained a short description of the skills an expert researcher applies when he or she is doing a literature search. The assignment was to design a blueprint for the training of this complex skill according to the 4C/ID-model on a blank answering form. After this test again a nine-point rating scale had to be completed as a measure of the mental effort spent on the transfer test. To be able to score the results of the transfer test a scoring form was developed consisting of twenty-eight *yes*/*no*-questions that checked to what extent and how accurately the strategy prescribed by the model had been applied in the transfer task. Every *yes* scored one point, and the sum score ranged from zero (no steps from the model taken) to 28 (all steps taken accurately). After the experiment, two independent raters scored the transfer tests using the form, showing an inter-rater agreement of .95. The average rating score was taken as the transfer score.

Procedure The experiment was carried out in eight sessions of about two hours, and in each session between one and seven students were tested simultaneously. These sessions took place in a multimedia lab that had seven computers connected to the Intranet of the Open University. Three computers had headphones attached to them. When the students entered the room they were randomly assigned to one of

the computers. Each computer showed a browser-window (without any of the menu options visible) set on a webpage displaying some general information about the experiment. When the students had finished reading, the experimenter told all the students to log in onto the actual instructions by typing in a password. All students started at the same time and studied the instructions all by themselves. The server on which the instructional website ran kept record of the mental effort scores of each participant.

After the instruction phase the three paper-and-pencil tests were administered. For each multiple-choice test the students got ten minutes, and for the transfer test they got thirty minutes.

Results

The variables under analysis were mental effort spent on the instructions, on the retention test and on the transfer test, and retention and transfer score. All scores were analysed with one-tailed t-tests. For all statistical tests, a significance level of .05 was applied. Table 1 shows the average scores on the dependent measures for the experimental groups.

Table 1: group means on dependent measures (standard deviations in brackets)

	audio	visual text
mental effort instructions	4.3 (0.8)	4.9 (0.9)
mental effort retention test	6.2 (0.8)	6.4 (1.2)
mental effort transfer test	6.4 (1.4)	7.1 (1.1)
retention score (0-50)	31.4 (6.1)	29.8 (5.4)
transfer score (0-28)	9.6 (6.2)	10.3 (5.4)

The reported mental effort during instructions showed a significant effect for the modality of text ($t(39) = -2.19$, $p < .05$). Students in the audio group had spent less effort than their colleagues in the visual group. The mental effort spent on the retention test showed no differences between the groups ($t(39) = -0.53$, $p > .10$). However, the mental effort scores in the transfer test did show a significant difference between the groups ($t(39) = 3.42$, $p < .05$), with the students in the audio group spending again less effort than their colleagues in the visual groups.

Although the audio group did a little better than the visual group on the retention test, this effect was not significant ($t(39) = 0.88$, $p > .10$). Also no significant difference was found between the groups on the transfer test ($t(39) = -0.40$, $p > .10$).

Discussion

The results of the first experiment show that the modality effect can be replicated with longer multimedia instructions on a non-technical subject like instructional design. Students in the audio group report lower mental effort scores during the instructions as a result of decreased memory load. This is confirmed by the fact that in both the retention and transfer test the students in the audio group score just as good as the students in the visual group. Moreover, getting the same result in the transfer test has cost them less mental effort. The modality effect is not as strong as in the experiments by Kalyuga et al. (1999) and Tindall-Ford et al. (1997), who found both lower mental effort and better test scores. However, the fact that students in the audio condition reach the same test results with less mental effort still points at the superiority of audio over visual text in multimedia instructions.

Experiment 2

In our second experiment we wanted to investigate the question if the modality effect can still be found if the students in the visual group get more time to relate the verbal information to the diagram. Therefore we not only varied the modality of the text, but also the pacing of the instructions. The system-paced groups were identical to the two groups in the first experiment, while students in the user-paced groups could set the pace of the instructions for themselves. This way we compared four groups: audio-user, audio-system, visual-user, and visual-system.

Method

Participants The participants were 81 second-year students from the Department of Education of the University of Gent in Belgium (age between 18 and 30 years; 8 males and 73 females). The experiment was part of a regular course on instructional design, but at the time of the experiment the students had not received any lessons on instructional design models yet. Eighteen participants were randomly assigned to the audio-user group, another 18 to the audio-system group, 24 to the visual-user group, and 21 to the visual-system group.

Materials The multimedia instructions were the same as in the first experiment, only two extra user-paced versions were created. In the audio-user version (Figure 2a), the students were able to replay the sentences they had just heard by clicking on a small *play*-button, while in the visual-user version (Figure 2b) students could reread the text as many times as they wanted to. To continue with the next piece of text students in both groups had to click on a forward arrow.

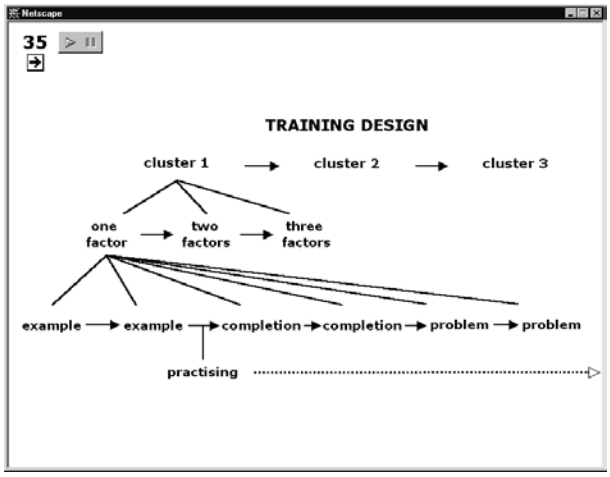


Figure 2a: Screen example of the audio-user version of the multimedia instructions.

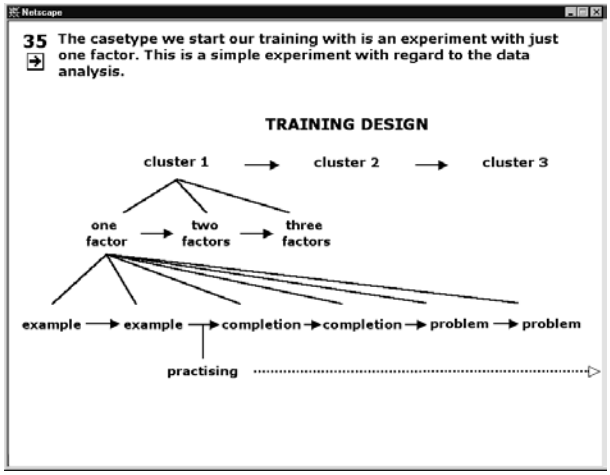


Figure 2b: Screen example of the visual-user version of the multimedia instructions.

The measurements were the same as in the first experiment, only this time all tests were presented on the computer. Moreover, the retention test consisted of 40 items taken from the retention test of the first experiment. The sum of the 40 items formed the total retention score (Cronbach's alpha = .68). After the experiment, two independent raters scored the transfer tests, showing an inter-rater agreement of .92.

Procedure The experiment was carried out in four sessions of about two-and-a-half hour, and in each session between fifteen and twenty-four students were tested simultaneously. These sessions took place in a classroom that had twenty-four multimedia computers connected to the Internet through the university network, with six computers for each experimental group. The procedure was almost identical to the first experiment. Only this time, students could immediately continue with the tests whenever they had finished studying the instructions. The server on which the instructional website ran kept record of the time spent on the learning task (in minutes), of the mental effort scores and of the retention score of each participant.

Results

The variables under analysis were training time, mental effort spent on the instructions, on the retention test and on the transfer test, and retention and transfer score. Except for training time, all scores were analysed with two-factor analyses of variance (ANOVAs), with modality (audio vs. visual text) and pacing of the instructions (system pacing vs. user pacing) as the between-subjects factors. For all statistical tests, a significance level of .05 was applied. Table 2 shows the average scores on the dependent measures for all four groups.

Table 2: group means on dependent measures (standard deviations in brackets)

	audio-user	audio-system	visual-user	visual-system
time on instructions	33.7 (3.3)		37.8 (5.5)	
mental effort instructions	4.3 (1.0)	4.1 (0.7)	4.0 (1.0)	4.2 (1.0)
mental effort retention test	6.5 (1.3)	6.7 (1.0)	6.5 (1.3)	6.7 (1.0)
mental effort transfer test	7.3 (1.1)	7.3 (1.5)	7.5 (1.1)	7.1 (1.5)
retention score (0-40)	26.5 (4.8)	28.9 (3.7)	28.6 (3.6)	25.3 (5.6)
transfer score (0-28)	17.8 (4.4)	17.7 (4.1)	16.8 (4.8)	14.1 (5.6)

With regard to the time spent on the instructions, only the two user-groups were compared, because in the system groups time was equal for all students (about 30 minutes). It showed that the students in the visual-user group had spent significantly more time on the instructions than the students in the audio-user group ($t(40) = -2.7$, $p < .01$, two-tailed).

There were no significant differences between the groups on mental effort during instructions. The same goes for the mental effort spent on the retention test, and for the mental effort spent on the transfer test.

The results on the retention test showed a significant interaction effect ($F(1,77) = 7.99$, $MSE = 20.16$, $p < .01$). In the two system groups, the audio group did better than visual text, while in the user groups this effect was reversed, with visual text outperforming the audio group. The scores on the transfer task showed a significant main effect for the modality of the text

($F(1,77) = 4.67$, $MSE = 23.07$, $p < .05$), with the students in the audio groups scoring higher than the students in the visual groups ($M = 17.8$, vs. $M = 15.5$, respectively). Inspection of the separate group means shows that especially the students in the visual-system group did worse that their colleagues in the audio groups. However, this interaction was statistically not significant.

Discussion

The results show that in the system-paced groups, a modality effect is found in terms of improved learning outcomes, but not in mental effort. This is a little different from the results in the first experiment in which the audio group spent less mental effort but did not have better test scores. This reversal might be accounted for by the fact that the second experiment was part of a regular course, and that the students in the audio condition were more prepared to invest the freed memory resources in the learning process itself, resulting in higher test scores with equal mental effort.

However, in the two groups in which the students set the pace of the instructions, no modality effect is found at all. Not only do the students in the visual-user group perform almost equally well on the transfer test, on the retention test they even outperform the students in the audio-user group. The visual-user group has taken more time to study the instructions, which confirms our idea that the modality effect in the system-paced condition is at least partly the result of a lack of time to relate the text to the diagrams in the visual-system group.

General Discussion

The results of both experiments show that replacing on-screen text with audio will only increase the effectiveness of multimedia instructions if the student has no control over the pacing of the instruction and the pace is set by the time of the narration. In that case we find either lower mental effort or better test results, even with a subject matter from a non-technical domain like instructional design. However, with more time (or the possibility to let the student determine the pace) visual-only instructions can be just as effective as bimodal instructions.

From a theoretical point of view, the results seem to indicate that the modality effect as demonstrated in earlier experiments can be accounted for in other terms than an increase in memory resources. One possible explanation is the lack of time to relate verbal to pictorial information in visual-only conditions. One of the things we will do in our future research is get a closer look at what actually happens when students are studying multimedia instructions by measuring eye-movements and look for different patterns in visual search.

Acknowledgements

This research project is partly funded by SPC Group, a multimedia company from ‘s Hertogenbosch, The Netherlands.

References

Baddeley, A. (1992). Working Memory. Science, 255, 556-559.

Jeung, H., Chandler, P., & Sweller, J. (1997). The role of visual indicators in dual sensory mode instruction. *Educational Psychology, 17*, 329-343.

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology, 13*, 351-371.

Mayer, R. E., & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology, 90*, 312-320.

Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia learning. *Journal of Educational Psychology, 91*, 358-368.

Mousavi, S. Y., Low, R., & Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology, 87*, 319-334.

Paas, F. G. W. C., & van Merriënboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review, 6*, 351-371.

Paas, F. G. W. C., van Merriënboer, J. J. G., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills, 79*, 419-430.

Park, I., & Hannafin, M. J. (1994). Empirically-based guidelines for the design of interactive multimedia. *Educational Technology, Research & Development, 41*, 66-85.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257-285.

Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251-296.

Tindall-Ford, S., Chandler, P., & Sweller, J. (1997). When two sensory modes are better than one. *Journal of Experimental Psychology: Applied, 3*, 257-287.

Van Merriënboer, J. J. G. (1997). *Training complex cognitive skills: A four-component instructional design model for technical training.* Englewood Cliffs, NJ: Educational Technology Publications.

Real World Constraints on the Mental Lexicon: Assimilation, the Speech Lexicon and the Information Structure of Spanish Words

Monica Tamariz (monica@ling.ed.ac.uk)
Department of Linguistics, AFB, 40 George Square
Edinburgh EH8 9LL, UK

Richard C. Shillcock (rcs@cogsci.ed.ac.uk)
Department of Cognitive Science, 2 Buccleuch Place
Edinburgh EH8 9LW, UK

Abstract

This paper focuses on the optimum use of representational space by words in speech and in the mental lexicon. In order to do this we draw the concept of entropy from information theory and use it to plot the information contour of words. We compare different representations of Spanish speech: a citation vs. a fast-speech transcription of a speech corpus and a dictionary lexicon vs. a speech lexicon. We also compare the information profiles yielded by the speech corpus vs. that of the speech lexicon in order to contrast the representation of words over two representational spaces: time and storage space in the brain. Finally we discuss the implications for the mental lexicon and interpret the analyses we present as evidence for a version of Butterworth's (1983) Full Listing Hypothesis.

Introduction

In this paper we focus on the optimum use of representational space by words over time (the sequence of sounds in speech) and over space (the storage site of the mental lexicon in the brain). We draw the concept of entropy from information theory and propose that it can be used to study the information structure of the set of words uttered in speech and of those stored in the mental lexicon in the face of the constraints of communication and of storage, respectively, in a potentially noisy medium.

We have two representational spaces for words: time and storage space. Further, we will consider the phonology and morphology of word systems. Our data sets are phonetic representations of words, and recent research demonstrates that information on the probabilistic distribution of phonemes in words is used in language processing (see Frisch, Large & Pisoni, 2000 for review). Morphology is involved in this research because we will be comparing groups of words with different inflectional and derivational features. We will initially assume the Full Listing Hypothesis (Butterworth, 1983): every word-form, including inflected and derived forms, is explicitly listed in the mental lexicon.

Shillcock, Hicks, Cairns, Chater and Levy (1995) suggest the general principle of the presentation of information in the brain that information should be spread as evenly as possible over time or over the representational space. Therefore, if the entropy of the mental lexicon is to be maximized so that the storage over a limited space is most efficient, then all the phonemes will tend to occur as evenly as possible in each segment position of the word. The phonology of each individual word, because it will have an effect on the entropy of the system, affects whether it is likely to become part of the mental lexicon.

Shillcock et al. stated that "the optimum contour across the phonological information in a spoken word is flat; fast-speech processes cause the information contour to become more level". We generalize this notion and propose the Levelling Effect of Realistic Representations (LERR): *processes that make the representation of words more accurate will flatten the information profiles*.

In order to test this, we will use Spanish word systems to calculate the slope and overall level of entropy of a citation (idealized pronunciation of the word in isolation) transcription and of a fast-speech (more realistic) transcription and of a dictionary lexicon and the speech lexicon. Our prediction is that the second system in each comparison should yield flatter information contours. We also compare a representation of words over time and another one over storage space - a speech corpus and the speech lexicon.

Entropy

We will use the concept of entropy in the context of information theory (Shannon, 1948), also employed in speech recognition studies (e.g. Yannakoudakis & Hutton, 1992). Entropy H is defined for a finite scheme

(i.e., a set of events such that one and only one must occur in each instance, together with the probability of them occurring) as a reasonable measure of the uncertainty or the information that each instance carries. E.g. the finite scheme formed by the possible outcomes when throwing a dice has maximum entropy: each side of the dice has 1/6 probability of occurring and it is very difficult to predict what the outcome will be. A loaded dice, on the other hand, has an unequal probability distribution, and the outcome is less uncertain. In this research, the possible events are the phonemes and allophones, and for each word only one of them can occur at each segment position.

For probabilities (p_1, p_2, p_3...p_n):

$H = - \Sigma (p_i \cdot \log p_i)$

The relative entropy H_{rel} is the measured entropy divided by the maximum entropy H_{max}, which is the entropy when the probabilities of each event occurring are equal and the uncertainty is maximized. Using H_{rel} allows us to compare entropies from systems with a different number of events (in this case, a system with 30 phonemes with another one with 50).

$H_{max} = \log n$

$H_{rel} = H / H_{max}$

Redundancy R is a measure of the constraints on the choices. When redundancy is high, the system is highly organized, and more predictable, i.e. some choices are more likely than others, as in the case of the loaded dice.

$R = 1 - H_{rel}$

In order to obtain the information profiles of words (see Figure 1), the entropy was calculated separately for each segment position in a set of left-justified words of equal length, i.e., for the first phoneme in the words, the second phoneme etc.

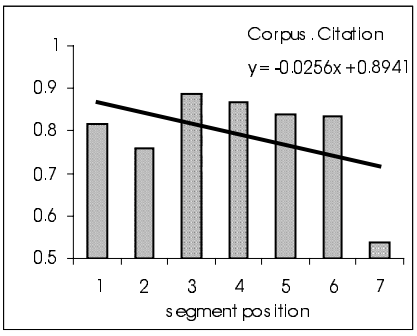


Figure 1: Information profile of 7-segment words from the citation transcription of the speech corpus.

The information profile of the word was measured as the linear trendline of these individual segment entropies. The slope (m) (multiplied by (-1)) of these trendlines and the mean relative entropy for each word length are shown in the figures below. E.g. In Figure 1, (-m)=0.0256. The flatness of the slope refers literally to how horizontal the trendline is.

Transcriptions

We have restricted ourselves to phonemic representations of word and will not report data concerning the distributions of phonemic features. We have used citation transcription rules (the idealised pronunciation of the isolated word) and fast-speech rules (an attempt to represent normal speech more realistically). Both citation and fast-speech rules were applied uniformly to the whole data sets. For the citation transcription we used 29 phonemes including 5 stressed vowels; for the fast-speech transcription we used 50 phonemes and allophones:

Citation transcription: Vowels: /a/, /e/, /i/, /o/, /u/, /á/, /é/, /í/, /ó/, /ú/. Consonants: /p/, /b/, /t/, /d/, /k/, /g/, /m/, /n/, /ɲ/, /ɾ/, /r/, /f/, /θ/, /s/, /ʝ/, /χ/, /l/, /ʎ/, /tʃ/.

Fast-speech transcription: The above plus semivowel /i̯/, /u̯/, voiced approximants /β̞/, /ð̞/, /ɣ̞/, voiceless approximants /β̞̥/, /ð̞̥/, /ɣ̞̥/, labiodental /m/, dental /n/ and /l/, palatalised /n/ and /l/, velarized /n/, /z/, dental voiced /s/, dental /s̪/, fricative /ɾ̞/, voiced /θ/ and a silenced consonant /Ø/. The transcription was made following the rules for consonant interactions, such as feature assimilation, set out by Rios Mestre (1999, chapter 5). Diphthongs were treated as two separate segments, as is usual in Spanish. Rules to mark stressed vowels were applied to all but monosyllabic words without an orthographic accent. For the corpus, the whole text was used, including repetitions and false starts of words. After deleting all the tags, the corpus was divided into chunks separated by pauses (change of speaker, comma, full stop, or pause marked in the transcription). The resulting text was transcribed automatically word by word (orthographic forms being replaced by phonetic forms) and then word boundary effects were introduced within the chunks, following the same rules as for the intra-word transcription.

Data

We used these three sets of data:

The speech corpus: a 707,000 word Spanish speech corpus, including repetitions and unfinished words. This corpus was developed by Marcos Marín of the Universidad Autonoma de Madrid in 1992 and contains transcribed speech from a wide range of registers and fields, from everyday conversation to academic talks and political speeches.

The dictionary lexicon: a 28,000 word Spanish word lexicon (the Spanish headword list of the Harrap Compact Spanish Dictionary, excluding abbreviations). This list does not include inflections, but approximately 40% of the words are derived words (we take the infinitive of verbs and the simple form of the noun as

the basic forms). This word system could represent a mental lexicon where that only word stems are listed and where inflected words are assembled during speech production.

The speech lexicon: the 42,000 word types found in the corpus. Some 80% of these types were derived and inflected words. We take this word system to be the most realistic representation of the mental lexicon, assuming Butterworth (1983)'s Full Listing Hypothesis, where all the wordforms are individually represented in the mental lexicon.

The dictionary lexicon and the speech lexicon share only ~30% of the words. The remaining ~70% of the words in the dictionary lexicon are mostly low frequency words which do not appear in our sample of speech. The new ~70% in the speech lexicon are verbal inflections (~35%), plurals and feminine inflections (~25%), some derived words absent from the dictionary lexicon (~4%), unfinished or mispronounced words (~4%) and proper nouns (~2%).

From these data, we used 4, 5, 6 and 7-segment transcriptions. Words were separated by length in order to see a clearer picture of the information profiles, especially as far as the word-ending contribution is concerned. Considering that the information profiles of Spanish words follows the same pattern as those of English words as seen in Shillcock et al. (1995), we can extend research in English to Spanish words. In English, word recognition typically occurs before the end of the word is uttered (Marslen-Wilson & Tyler, 1980), and information about word-length is typically available once the nucleus is being processed (Grosjean, 1985). It is, therefore, legitimate to assume that recognition processes are restricting their activities to the subset of words in the lexicon that match the word being uttered both in terms of initial segments and approximate overall length. The particular word lengths were chosen because the structure of shorter words is simpler, and the effects are less likely to be obscured by greater variation in the internal structure of each word-length group. These word lengths are equidistant from the modes of the word-length distribution of the three data sets (lexicon: mode = 8, speech lexicon: mode = 7 and speech corpus: modes = 2, 4 – the mode of the normal distribution is 4, but the proportion of 2-segment words is even higher, accounting for 32% of all tokens). The sum of these four word lengths accounts for 41% of the dictionary lexicon, 45% of the speech lexicon and 37% of the speech corpus.

The effect of the transcription

Shillcock et al. (1995) showed that fast-speech processes cause the information contour to become more level for English, German, Welsh, Irish and Portuguese. Here we compare the slope of the information profiles of 4-7 segment words from the corpus transcribed with citation rules and with fast-speech rules.

As predicted by the LERR principle, Figure 2 confirms that this is also the case for Spanish. The information profile is consistently flatter for the more realistic fast-speech transcriptions in all word lengths. Note that in the figure, a higher value of (–m) indicates a steeper profile.

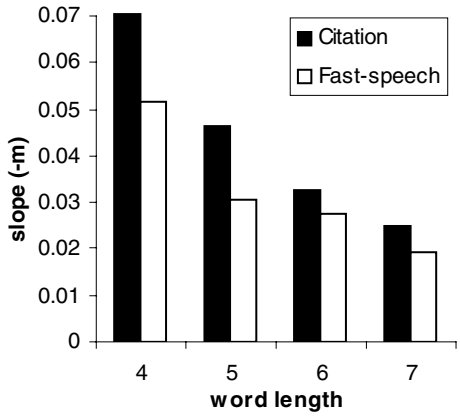


Figure 2: Slopes of the information profiles of the citation and the fast-speech transcriptions applied to the corpus, over the four word lengths.

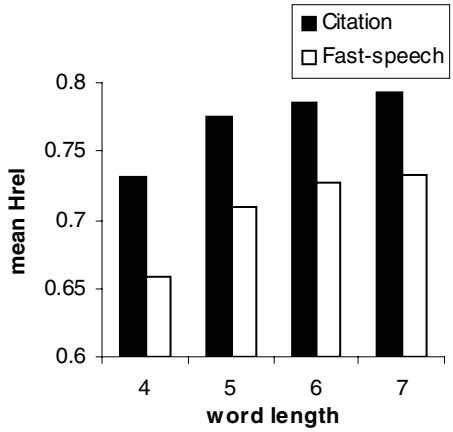


Figure 3: Mean relative entropy of the citation and fast-speech transcriptions over the four word lengths.

Figure 3 shows how the overall entropy is lower for the fast-speech transcription: when we introduce the allophones and the assimilation rules, the system becomes more redundant and thus, more predictable.

The Speech Lexicon

Some current models of lexical access propose two parallel word recognition routes, a whole-word route and a morpheme-based one (e.g. Wurm (1997) for English; Colé, Segui & Taft (1997) for French; Laine, Vainio & Hyona (1999) for Finnish). Following this hypothesis, the full forms of words need to be stored in the mental lexicon (cf. Butterworth, 1983). It is relevant, then, to study the behaviour of the set of all word types, including derived and inflected words, that appear in speech: the speech lexicon.

We have seen that fast-speech transcriptions yield flatter information contours than citation transcriptions, so we will use the fast-speech transcriptions of the speech lexicon, the lexicon and the corpus.

Comparing the slopes of the information profiles of the speech lexicon on the one hand and the dictionary lexicon and the corpus on the other hand will help characterize the active mental lexicon.

Speech lexicon vs. dictionary lexicon

The speech lexicon contains inflected and derived forms, and does not contain the more obscure words that can be found in the dictionary. The LERR principle that data that are closer to real speech should produce flatter information contours is confirmed in Figure 4, where we see that the values of the slope of the information profile of the speech lexicon are lower than those of the dictionary lexicon.

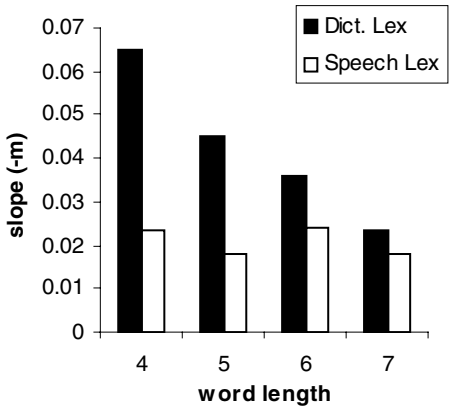


Figure 4: Slopes of the information profiles of the dictionary lexicon and the speech lexicon over the four word lengths.

Figure 5 shows that the overall entropy level is higher for the speech lexicon. This means that the speech lexicon is less redundant than the dictionary lexicon. The representational space is now a limited amount of memory storage space in the brain, and for maximal efficiency redundancy has to be reduced as much as possible. The results from both the slopes and the entropy levels support the Full Listing Hypothesis that all wordforms, particularly inflected forms, are listed in the mental lexicon – the system that includes all wordforms (the speech lexicon) could be stored more efficiently over a limited representational space.

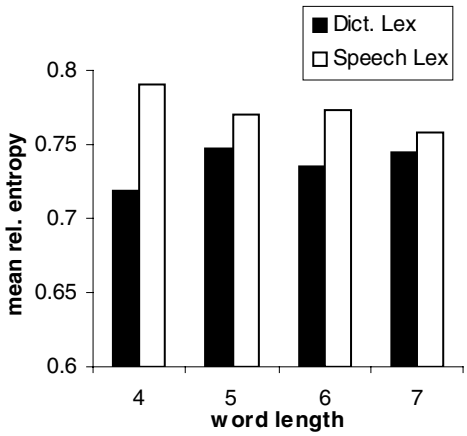


Figure 5: Mean relative entropy of the dictionary lexicon and the speech lexicon over the four word lengths.

Speech lexicon vs. corpus

The fact that entropy and redundancy statistics obtained from a lexicon are different from those obtained from a corpus has been noted by Yannakoudakis and Angelidakis (1988). Here we are comparing the word tokens with the word types in a speech corpus. Figures 6 and 7 show that the speech lexicon has consistently flatter slopes and higher entropy levels than the corpus.

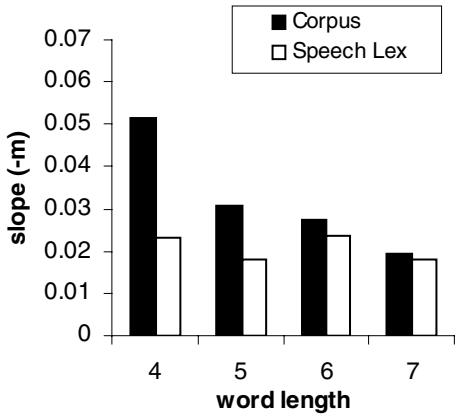


Figure 6: Slopes of the information profiles of the corpus and the speech lexicon across the four word lengths.

We are comparing two representational spaces: words in the brain are constrained by a limited space and words uttered over time are constrained by the efficiency of communication. We saw in the last section that the flat slopes and high entropy levels of the speech lexicon information profiles are best suited to enhance storage efficiency. Slopes in the corpus are relatively flat, but still steeper than those of the speech lexicon. This may reflect the fact that there are other factors affecting the information contour of words in speech, such the need to encode cues to lexical segmentation (signals that indicate where words begin and end). These other factors may be interacting with the optimization of communication.

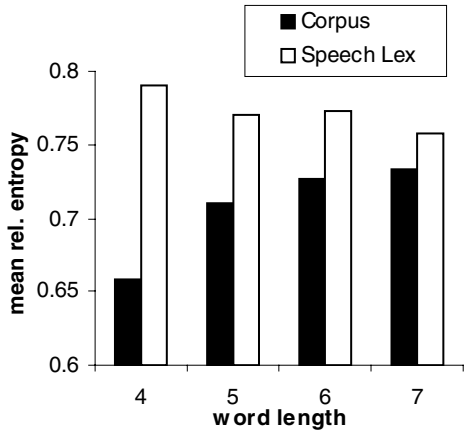


Figure 7: Mean relative entropy of the corpus and the speech lexicon across the four word lengths.

The corpus presents lower entropy levels than the speech lexicon. Speech over time is not constrained by space limitations, but rather by the need to communicate efficiently. The higher redundancy means that this system reduces the uncertainty and is indeed better for communication.

Discussion

The present study points in the direction of the LERR principle that the more realistic data - the fast-speech transcription and the speech lexicon - produce flatter information profiles.

The flatter profile of the fast-speech transcription can be partly explained in terms of the Markedness Ordering Principle (Shillcock et al., 1995) that when consonant interactions introduce phonological ambiguity, the ambiguity introduced is always in the direction of a less frequent phoneme. As for the comparison between lexicons, let us remember that the 70% of words in the speech lexicon that do not appear in the dictionary lexicon are mostly inflected words, and the 70% of words in the dictionary lexicon not present in the speech lexicon are mainly low-frequency words. The flatter profile of the speech lexicon is due to the fact that the inflected words (which are derived from one third of the dictionary lexicon words) yield a flatter profile than the low-frequency dominated group. This suggests that inflected words are included in the mental lexicon, and so it supports the Full Listing Hypothesis.

Additionally, the overall level of entropy and redundancy gives us an insight into the degree of complexity of a system. Highly organized systems will show low entropy and high redundancy. Fast-speech rules make the system more redundant than the citation rules. This higher predictability helps to deal with the loss of information produced by noise and thus enhance communication. The speech lexicon is less redundant than the dictionary lexicon. Here again, the higher entropy must be attributable to the fact that the phonemes in inflected forms are more evenly distributed over the phonological space than the more obscure words present in the dictionary lexicon.

The comparison between the corpus and the speech lexicon shows the features of the representation that has evolved to enhance communication and storage, respectively. Both systems are “realistic”, and indeed both show relatively flat information contours, but more so the speech lexicon, suggesting that communication has other constraints that interact with this measure, such as word-boundary recognition. This is true particularly for shorter words. The fact that the corpus is markedly more redundant than the speech lexicon is only to be expected, since it reflects the added complexity of different word-frequencies.

In conclusion, we have shown that it is possible to use psychological theories of the mental lexicon and spoken word recognition to make testable predictions concerning distributional information in large samples of language, and, conversely, that data from information distribution may potentially falsify particular aspects of those psychological theories. Our current conclusions from the analyses of Spanish favour versions of Butterworth's original Full Listing Hypothesis, in which all the wordforms encountered in speech are individually stored.

Acknowledgments

This research has benefited from the support of EPSRC studentship award nr. 00304518.

References

Butterworth, B. (1983). Lexical representation. In B. Butterworth (Ed.), *Development, writing and other language processes, Vol. 2*, London: Academic Press.

Colé, P., Segui, J., & Taft, M. (1997, Words and morphemes as units for lexical access, *Journal of memory and language*, *37 (3)*, 312-330.

Frisch, S. A., Large, N. R. & Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords, *Journal of Memory and Language*, *42 (3)*, 481-496.

Grosjean, F. (1985). The recognition of words after their acoustic offset: Evidence and implications. *Perception and Psychophysics*, *38*, 299-310.

Laine M., Vainio, S., & Hyona, J. (1999). Lexical access routes to nouns in a morphologically rich language, *Journal of memory and language*, *40 (1)*, 109-135.

Marcos Marin, F. (1992*). Corpus oral de referencia del español*, Madrid: UAM.

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding, *Cognition*, *8*, 1-71.

Ríos Mestre, A. (1999). La transcripción fonética automática del diccionario electrónico de formas simples flexivas del español: estudio fonológico en el léxico, *Estudios de Lingüística Española*, *4*.

Shannon, C. E. (1948). A mathematical theory of communication, *Bell System Technology Journal*, *27* (July), 379-423 and (October), 623-656.

Shillcock, R.C., Hicks, J., Cairns, P., Chater, N., & Levy, J. P. (1995). Phonological reduction, assimilation, intra-word information structure, and the evolution of the lexicon of English: Why fast speech isn't confusing. *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 233-238), Hillsdale, NJ: Lawrence Erlbaum Associates.

Yannakoudakis, E. J. & Hutton, P. J. (1992). An assessment of N-phoneme statistics in phoneme guessing algorithms which aim to incorporate phonotactic constraints, *Speech communcation*, *11*, 581-602.

Yannakoudakis, E. J. & Angelidakis, G. (1988). An insight into the entropy and redundancy of the English dictionary, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *10 (6)*, 960-970.

Wurm L. H. (1997). Auditory processing of prefixed English words is both continuous and decompositional, *Journal of memory and language*, *37 (3)*, 438-461.

The Rational Basis of Representativeness

Joshua B. Tenenbaum & Thomas L. Griffiths
Department of Psychology
Stanford University
Stanford, CA 94305-2130 USA
jbt,gruffydd @psych.stanford.edu

Abstract

Representativeness is a central explanatory construct in cognitive science but suffers from the lack of a principled theoretical account. Here we present a formal definition of one sense of representativeness – what it means to be a good example of a process or category in the context of Bayesian inference. This analysis clarifies the relation between representativeness as an intuitive statistical heuristic and normative principles of inductive inference. It also leads to strong quantitative predictions about people's judgments, which compare favorably to alternative accounts based on likelihood or similarity when evaluated on data from two experiments.

Why do people think that Linda, the politically active, single, outspoken, and very bright 31-year-old, is more likely to be a feminist bankteller than to be a bankteller, even though this is logically impossible? Why do we think that the sequence is more likely than the sequence to be produced by flipping a fair coin, even though both are equally likely? The standard answer in cognitive psychology (Kahneman & Tversky, 1972) is that our brains are designed to judge "representativeness", not probability: Linda is more representative of feminist banktellers than of banktellers, and is more representative of flipping a fair coin than is , despite anything that probability theory tells us.

Not only errors in probabilistic reasoning, but numerous other phenomena of categorization, comparison, and inference have been attributed to the influence of representativeness (or prototypicality or "goodness of example"; Mervis & Rosch, 1981; Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Rips, 1975). However, a principled account of representativeness has not been easy to come by. Its leading proponents (Kahneman & Tversky, 1996; Mervis & Rosch, 1981) have asserted that representativeness should be defined only operationally in terms of people's judgments; an a priori, analytic definition need not be given. Critics have countered that this concept is too vague to serve as an explanation of intuitive probability judgment (Gigerenzer, 1996).

This paper presents a framework for constructing rational models of representativeness, based on a Bayesian analysis of what makes an observation a good example of a category or process. The goal is to identify precisely one sense of representativeness and show that it has a rational basis in normative principles of inductive reasoning. We will first point out some shortcomings of previous accounts based on likelihood or similarity, and show how a Bayesian approach can overcome those problems. We will then compare the quantitative predictions of Bayesian, likelihood, and similarity models on two sets of representativeness judgments.

Previous approaches

Likelihood. In trying to relate intuitions about representativeness to rational statistical inferences, a natural starting point is the concept of likelihood. Let d denote some observed data, such as a sequence of coin tosses, and h denote some hypothesis about the source of d, such as flipping a fair coin. The probability of observing d given that h is true, $P(d\ h)$, is called a likelihood. Let $R(d,h)$ denote representativeness – how representative the observation d is of the generative process in h.

Gigerenzer & Hoffrage (1995) have proposed that representativeness, to the extent that it can be defined rigorously, is equivalent to likelihood: $R(d,h) = P(d\ h)$. This proposal is appealing in that, other factors aside, the more frequently h leads to observing d, the more representative d should be of h. It is also consistent with some classic errors in probability judgment, such as the conjunction fallacy: a person is almost certainly more likely to match Linda's description given that she is a bankteller and a feminist than given only that she is a bankteller.

While likelihood and representativeness seem related, however, they are not equivalent. Two observations with equal likelihood may differ in representativeness. Knowing that and are equally likely to be produced by a fair coin does not change our judgment that the latter is the more representative outcome. Tversky & Kahneman (1983) provide several examples of cases in which a more representative outcome is actually less likely. Any sequence of fair coin flips, such as , is less likely than one of its subseqences, such as or , but may easily be more representative. More colorfully, "being divorced four times" is more representative of Hollywood actresses than is "voting democratic", but the former is certainly less likely.

Figure 1 illustrates a simple version of the dissociation between representativeness and likelihood. Each panel shows a sample of three points from a Gaussian distribution. With independent sampling, the total likelihood of a sample equals the product of the likelihoods for

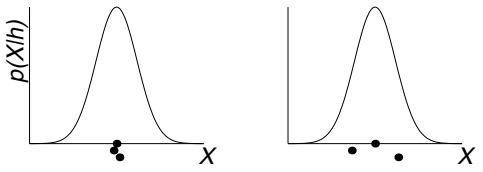


Figure 1: Given a normal distribution, the left sample has greater likelihood but the right is more representative.

each item in the sample. Thus the left sample has much greater likelihood, because each point is much closer to the peak of the distribution than in the right sample. Yet the more spread-out sample on the right seems more representative. We tested this intuition in a survey of 138 Stanford undergraduates. They were first shown a normally distributed set of thirty "widgets" produced by a factory. The widgets were simple drawings resembling nuts or bolts, varying only in their sizes. They were then shown three different samples, each with three widgets, and asked to rate on a scale of 1-10 how representative each sample was of the widgets produced by this factory. Each sample contained a point at the mean of the original distribution, and points at $z = \ 2.85$ ("broad sample"), $z = \ 1$ ("intermediate sample"), or $z = \ 0.05$ ("narrow sample"). The intermediate sample, with a standard deviation similar to the population, received a significantly higher rating than did the much more likely narrow sample (7.1 vs. 5.2, $p < .05$). The broad sample, with lowest likelihood of all, also received a lower rating (6.9) than the intermediate sample, but not by a significant margin.

We also tested whether intermediate-range samples are more representative for natural categories, using as stimuli black-and-white pictures of birds. In a design parallel to the widget study, 135 different Stanford undergraduates saw three samples of birds, each containing three members, and rated how representative they were of birds in general. The samples consisted of either three robins ("narrow"); a robin, an eagle, and a seagull ("intermediate"); or a robin, an ostrich, and a penguin ("broad"). Although the robins were individually rated as more representative than the other birds (by a separate group of 100 subjects), the set of three robins was considered the least representative of the three samples. As with the widgets, the intermediate sample was rated more representative (6.3) than either the narrow (5.1) or broad (5.3) samples ($p < .05$ for both differences).

For natural categories as well as for the artificial widgets, a set of representative examples turns out not to be the most representative set of examples. Sample likelihood, because it is merely the product of each example's individual likelihood, cannot capture this phenomenon. At best, then, likelihood may be only one factor contributing to the computation of representativeness.

Similarity. Most attempts to explicate the mechanisms of representativeness, including that of Kahneman & Tversky (1972), rely not on likelihood but on some sense of similarity. That is, an observation d is representative of a category or process h to the extent that it is similar to the set of observations h typically generates.

Similarity seems to avoid some of the problems that likelihood encounters. may be more representative of a fair coin than because it is more similar on average to other coin flip sequences, based on such features as the number of heads or the number of alternations. Likewise, someone who has been divorced four times may be more similar to the prototypical Hollywood actress than someone who votes democratic, if marital status is weighted more heavily than political affiliation in computing similarity to Hollywood actresses.

However, the explanatory power of a similarity-based account hinges on being able to specify what makes two stimuli more or less similar, what the relevant features are and how are they weighted. Similarity unconstrained is liable to lead to circular explanations: having had multiple divorces is more representative of Hollywood actresses because marital status is more highly weighted in computing similarity to Hollywood actresses, but why is marital status so highly weighted, if not because having multiple divorces is so typical of Hollywood actresses?

Equating representativeness with similarity also runs into a problem when evaluating the representativeness of a set of objects, as in Figure 1. Similarity is usually defined as a relation between pairs of stimuli, but here we require a judgment of similarity between two sets of stimuli, the sample and the population. It is not immediately obvious how best to extend similarity from a pairwise to a setwise measure. The individual elements of the left sample are certainly more similar to the average member of the population than are the elements of the right sample. The left sample also comes closer to minimizing the average distance between elements of the population and elements of the sample. If similarity between sets is defined according to one of these measures, it will fail to match up with representativeness.

Finally, and most problematic for our purposes here, a definition in terms of similarity fails to elucidate the rational basis of representativeness, and thus brings us no closer to explaining when and why representativeness leads to reasonable statistical inferences. Hence we seem to be left with two less-than-perfect options for defining representativeness: the simple, rational, but clearly insufficient concept of likelihood, or the more flexible but notoriously slippery concept of similarity.

A Bayesian analysis

In this section we present a Bayesian analysis of representativeness that addresses some of the shortcomings of the likelihood and similarity proposals. As with likelihood, Bayesian representativeness takes the form of a simple probabilistic quantity, which in fact includes likelihood as one component. But like the similarity approach, it can account for dissociations of representativeness and likelihood, when a less probable feature of the stimuli is also more diagnostic of the process or category in question. Moreover, it applies just as well to evaluat-

ing the representativenes of a set of examples (e.g. Figure 1) as it does to individual examples.

Our notion of a "good example" is defined in the context of a Bayesian inductive inference task. As above, let d denote some observed data, and let $H = h_1, \ldots h_n$ denote a set of n hypotheses (assumed to be mutually exclusive and exhaustive) that might explain the observed data. For each h_i, we require both the likelihood $P(d\ h_i)$ and a prior probability, $P(h_i)$, which expresses the degree of belief in h_i before d is observed. Let $\bar{h}_i = h_j \quad H : j = i$ denote the negation of hypothesis h_i, the assertion that some hypothesis other than h_i is the true source of d. Then we define our measure of representativeness $R(d, h_i)$ to be the logarithm of the likelihood ratio

$$L(d\ h_i) = \frac{P(d\ h_i)}{P(d\ \bar{h}_i)}. \tag{1}$$

This definition is motivated by Bayes' rule, which prescribes a degree of belief in hypothesis h_i after observing d given by the posterior probability

$$P(h_i\ d) = \frac{P(d\ h_i)P(h_i)}{P(d)}. \tag{2}$$

Defining the posterior odds $O(h_i\ d) = P(h_i\ d)\ \ P(\bar{h}_i\ d) = P(h_i\ d)\ \ (1 - P(h_i\ d))$, and the prior odds $O(h_i) = P(h_i)\ \ (1 - P(h_i))$, we can write Bayes' rule in the form:

$$\log O(h_i\ d) = \log L(d\ h_i) \quad \log O(h_i). \tag{3}$$

Equation 3 shows why the log likelihood ratio, $\log L(d\ h_i)$, provides a natural measure of how good an example d is of h_i: it indicates the extent to which observing d increases or decreases the posterior odds of h_i relative to the prior odds. Researchers in statistics (Good, 1950), artificial intelligence (Pearl, 1988), and philosophy of science (Fitelson, 2000) have previously considered $\log L(d\ h_i)$ as the best measure for the weight of evidence that d provides for h_i, because it captures the unique contribution that d makes to our belief in h_i independently of all other knowledge that we have (reflected in $P(h_i)$).

To compute $R(d, h_i)$ in the presence of more than one alternative hypothesis, we express it in the form

$$R(d, h_i) = \log \frac{P(d\ h_i)}{\sum_{h_j\ H} P(d\ h_j)P(h_j\ \bar{h}_i)}. \tag{4}$$

$P(h_j\ \bar{h}_i)$ is the prior probability of h_j given that h_i is not the true explanation of d: 0 when $i = j$ and $P(h_j)\ \ (1 - P(h_i))$ otherwise. Equation 4 shows that d is representative of h_i to the extent that its likelihood under h_i exceeds its average likelihood under alternative hypotheses.

To illustrate the analysis concretely, consider the simple case of two coinflip sequences, and . Unlike the likelihood model, we cannot compute how representative an observation is of a hypothesis without specifying the alternative hypotheses that an observer might consider. In the interests of simplicity, we consider just three relevant hypotheses about the origins of and : a fair coin (h_F), a two-headed coin (h_T), and a weighted coin (h_W) that comes up heads with probability 3/5. The likelihoods of the two sequences under these hypotheses are, for the fair coin, $P(\quad h_F) = P(\quad h_F) = (1\ \ 2)^5 = 0.03125$; for the two-headed coin, $P(\quad h_T) = 1$ while $P(\quad h_T) = 0$; and for the weighted coin, $P(\quad h_W) = (3\ \ 5)^5 = 0.0778$ while $P(\quad h_W) = (3\ \ 5)^3(2\ \ 5)^2 = 0.0346$. For concreteness, we choose specific prior probabilities for these hypotheses: $P(h_F) = 0.9$, $P(h_T) = 0.05$, and $P(h_W) = 0.05$. Substituting these numbers into Equation 4, we have $R(\quad , h_F) = \log \frac{0.03125}{1\times 0.05\ \ 0.1\ \ 0.0778\times 0.05\ \ 0.1} = -2.85$, while $R(\quad , h_F) = \log \frac{0.03125}{0\times 0.05\ \ 0.1\ \ 0.0346\times 0.05\ \ 0.1} = 0.59$. This result, that is more representative of a fair coin than , accords with intuition and holds regardless of the prior probabilities we assign to the three alternative hypotheses. In a later section, we go beyond a qualitative reconstruction of intuitions to test a quantitative model of representativeness judgments for sequences of coin flips.

The Bayesian approach also accounts for cases where a sample with lower likelihood appears more representative. For instance, $P(\quad h_F)$ is strictly lower than either $P(\quad h_F)$ or $P(\quad h_F)$, but is no less representative than . The Bayesian account also offers an intuitively compelling definition of representativeness for a set of examples, such as the widgets in Figure 1. We demonstrate by computing the representativeness for a sample X from a Gaussian population h_1. Let $x_1, \ldots, x_N$ be the N examples in X, m be the mean of X, and $S = \sum_i (x_i - m)^2$ the sum-of-squares. Let h_1 have mean μ and variance σ^2. We take the hypothesis space H to include all possible Gaussian distributions in one dimension – each a conceivable alternate explanation for the sample X. Because H is an uncountably infinite set, the sum in the denominator of Equation 4 becomes an integral. Assuming an uninformative Jeffreys prior on μ, σ (Equation 3 of Minka, 1998), our expression for Bayesian representativeness in Equation 4 then reduces to

$$R(X, h_1) = N \log S - \frac{1}{\sigma^2}\left[N(m - \mu)^2 \quad S\right], \tag{5}$$

plus a term that depends only on N and σ^2.

Equation 5 is maximized when $m = \mu$ and $S\ N = \sigma^2$, that is, when the mean and variance of the sample X match the mean and variance of the population h_1. This result is intuitive, and it accounts for why people preferred intermediate samples of widgets or birds over broad or narrow samples in the surveys described above: the $N \log S$ term penalizes narrower samples and the $-S\ \sigma^2$ penalizes broader samples. Yet this result is also not particularly surprising. More interestingly, Equation 5 gives a general metric for scoring the representativeness of any sample from a Gaussian distribution, which we will test quantitatively against people's judgments in the following section.

Quantitative modeling

In this section, we present quantitative models of representative judgments for two kinds of stimuli: sequences of coin flips and sets of animals. For each data set, we compare the predictions of Bayesian, likelihood-based, and similarity-based models.

Coin flips

Methods. 278 Stanford undergraduates rated the representativeness of four different coin flip sequences for each of four hypothetical generative processes, under the cover story of helping a casino debug a new line of gambling machines. The sequences were $d_1 =$, $d_2 =$, $d_3 =$, and $d_4 =$. The generative processes were h_1 = "A fair coin", h_2 = "A coin that always alternates heads and tails", h_3 = "A coin that mostly comes up heads", and h_4 = "A coin that always comes up heads". The orders of both sequences and hypotheses were randomized across subjects. Representativeness judgments were made on a scale of 1-7.

Bayesian model. While people could construct an arbitrarily large hypothesis space for this task, we make the simplifying assumption that their hypothesis space can be approximated by just the four hypotheses that they are asked to make judgments about. We constructed simple probabilistic models for each hypothesis h_i to generate the necessary likelihoods $P(d_j\ h_i)$. Priors for all hypotheses were assumed to be equal. To model h_1, "a fair coin", all likelihoods were set equal to their true values of 1 2^8. To model h_3, "mostly heads", and h_4, "always heads", we used binomial distributions with $p = 0.85$ and $p = 0.99$, respectively. In some sense, these p values represent free parameters of the model, but their values are strongly constrained by the meaning of the words "mostly" and "always". Their exact values are not crucial to the model's performance, as long as "always" is taken to mean something like "almost but not quite always" (i.e. $p < 1.0$). To model h_4, "always alternates heads and tails", we used a binomial distribution over the seven possible state transitions in each sequence, again with "always" translated into probability as $p = 0.99$. All model predictions were then given by Equation 4.

Likelihood model. This model treats representativeness judgments simply as $P(d_j\ h_i)$, as specified above.

Similarity model. We defined a simple similarity-based model in terms of two intuitively relevant features for comparing sequences: the number of heads in each sequence and the number of alternations in each sequence. Let α_j be the number of heads in sequence j, and β_j be the number of alternations. Then the similarity of sequences d_i and d_j is defined to be

$$sim(d_i, d_j) = \exp\left(-w_\alpha\ \alpha_i - \alpha_j\ - w_\beta\ \beta_i - \beta_j\ \right), \quad (6)$$

where w_α and w_β are the weights given to these two features. To compute similarity between a sequence and a generating hypothesis, we construct a prototype for each hypothesis based on the mean values of α and β over the whole distribution of sequences generated by that hypothesis. For example, for h_2, $\alpha = 4$ and $\beta = 7$; for h_3 (again assuming "mostly" means with probability 0.85), α 6.8 and β 1.8. Lastly, we define the representativeness of sequence i for hypothesis j as $R(d_i, h_j) = sim(d_i, h_j)\ \sum_k sim(d_i, h_k)$. The dimensional weights w_α and w_β are free parameters optimized to fit the data, giving $w_\alpha = 1$, $w_\beta = 0.4$.

Results. To compensate for nonlinear transformations that might affect the 1-7 rating scale used by subjects, the predictions of each model were first transformed according to a power function with a power γ chosen to optimize each model's fit, and then mapped onto the same interval spanned by the data. This gives both the likelihood model and the Bayesian model one free parameter plus two constrained parameters (corresponding to the meanings of "mostly" and "always"), while the similarity model has three free parameters (w_α, w_β, and γ) and the same two constrained parameters. All three models correlate highly with subjects' representativeness judgments, although the Bayesian model has a slight edge with $r = 0.94$, versus 0.87 for the likelihood model and 0.92 for the similarity model. Figure 2 presents an item-by-item analysis, showing that the Bayesian model captures virtually all of the salient patterns in the data.

Animals

Methods. We used data reported by Osherson, Smith et al. (1990; Tables 3 and 4) in a study of category-based induction. They asked one group of subjects to judge pairwise similarities for a set of 10 mammals, and a second group of subjects to judge the strengths of 45 arguments of the form x_1 has property P, x_2 has property P, x_3 has property P, therefore all mammals have property P , where x_1, x_2 and x_3 are three different kinds of mammals and P is a blank biological predicate. Such judgments of argument strength are not the same thing as judgments of representativeness, but for now we take them as a reasonable proxy for how representative the sample $X = \ x_1, x_2, x_3\ $ is of the set of all mammals.

Bayesian model. We assume that people's hypothesis space includes the category of all mammals (h_M), as well as an infinite number of alternative hypotheses. For simplicity, we model all hypotheses as Gaussian distributions in a two-dimensional feature space obtained from a multidimensional scaling (MDS) analysis of the similarity judgments in Osherson et al. (1990). This allows us to apply essentially the same analysis used in the previous section to compute the representativeness of a sample from a Gaussian distribution (Equation 5), and also parallels the original approach to modeling category-based induction of Rips (1975). The MDS space for animals is shown in Figure 3. The large gray oval indicates the one-standard-deviation contour line of h_M, which we take to be the best fitting Gaussian distribution for the set of all ten mammals. We assume the set H of alternative hypotheses includes all Gaussians in this two-dimensions

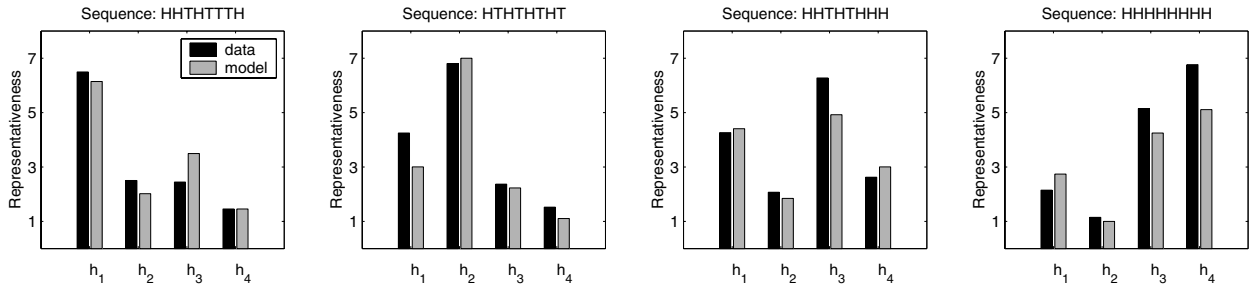


Figure 2: Representativeness judgments for coin flip sequences. Each panel shows subjects' mean judgments and the Bayesian model predictions for the representativeness of one sequence with respect to four different generating hypotheses: h_1 = "A fair coin", h_2 = "A coin that always alternates heads and tails", h_3 = "A coin that mostly comes up heads", and h_4 = "A coin that always comes up heads".

space, and we again use the uninformative Jeffreys' prior $P(h)$ (Minka, 1998; Equation 3). How representative a sample X (e.g. horse, cow, squirrel) is of all mammals can then be computed from a multidimensional version of Equation 5 (ignoring terms equal for all samples):

$$R(X,h_m) = N\log|\mathbf{S}| - N(\mathbf{m}-\mu)^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{m}-\mu) - \mathrm{trace}(\mathbf{S}\mathbf{V}^{-1}), \quad (7)$$

where m is the mean of X, $\mathbf{S} = \sum_i(\mathbf{x}_i-\mathbf{m})^{\mathrm{T}}(\mathbf{x}_i-\mathbf{m})$, $\mathbf{x}_i$ are the MDS coordinates of example i, N is the number of examples in X, and μ and $\mathbf{V}$ are the mean and covariance matrix of h_M (Minka, 1998). Equation 7 measures the representativeness of any sample X of N mammals in terms of the distance between the best fitting Gaussian for the sample (mean $\mathbf{m}$, covariance $\mathbf{S}$/N) and the best fitting Gaussian for the set of all mammals (mean μ, covariance $\mathbf{V}$). Figure 3 illustrates this graphically, by plotting one-standard-deviation contours for three samples that vary in how representative they are of the set of all mammals. Observe that the more representative the sample, the greater the overlap between its best-fitting Gaussian and the best-fitting Gaussian for the whole set.

Similarity-based models. Osherson et al. (1990) report pairwise similarity judgments for the animals, but to construct a similarity-based model of this representativeness task, we need to define a setwise measure of similarity between any sample of three animals and the set of all mammals. The similarity-coverage model proposed by Osherson et al. defines this quantity as the sum of each category instance's maximal similarity to the sample: $R(X,h_M) = \sum_j \max_i sim(i,j)$, where j ranges over all mammals and i ranges over just those in the sample X. A more traditional similarity-based model might replace the maximum with a sum: $R(X,h_M) = \sum_j\sum_i sim(i,j)$. Osherson et al. (1990) consider both max-similarity and sum-similarity models but favor the former as it is more consistent with their phenomena. However, there seems to be little a priori reason to prefer max-similarity, and indeed most similarity-based models of classification are closer to sum-similarity, so we consider both here.

Other models. We also compare the predictions of a simple likelihood model, which equates representativeness with $P(X\ h_M)$, and Sloman's (1993) feature-based model. Heit (1998) also presented a Bayesian model of category-based induction tasks, but because his model depends heavily on the choice of priors, it does not make strong quantitative predictions that can be evaluated here.

Results. Figure 3 plots the argument strength judgments for 45 arguments versus the representativeness predictions of the probabilistic and similarity-based models. Both the Bayesian and max-similarity models predict the data reasonably well ($r = 0.80$ vs. $r = 0.88$), with no significant difference between them ($p > .2$). Neither of these models has any free numerical parameters. With one free parameter, the feature-based model performs slightly worse ($r = 0.71$). Interestingly, both the likelihood and sum-similarity models show a weak *negative* correlation with the data ($r = -.31$, $r = -.26$). This discrepancy directly embodies the insight of Figure 1: high likelihood can yield low representativeness when the sample is tightly clustered near the mean, as in the sample of horse, cow, rhino (ellipse **C** in Figure 3). Sum-similarity performs as poorly as likelihood because it is essentially a nonparametric estimate of likelihood; likewise, max-similarity performs well because it correlates highly with Bayesian representativeness.

Discussion

Overall, the Bayesian models provide the most satisfying account of these two data sets. On the coinflip data, not only does Bayes obtain the highest correlation, but it does so with the minimal number of free parameters. On the animals data, Bayes obtains a correlation competitive with the best of the other models, max-similarity, even though it is based on less than half as much input data (20 MDS coordinates versus 45 raw similarity judgments) and may be hindered by information lost in the MDS preprocessing step. Most importantly, the Bayesian models are based on a rational analysis, which provides a single principled definition of representativeness applicable across the two quite different domains of coinflips and

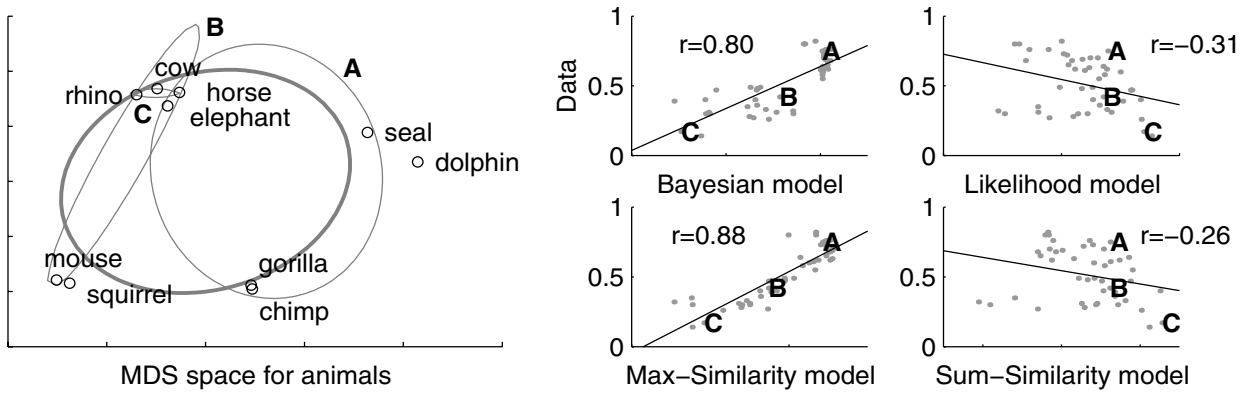


Figure 3: Modeling representativeness for sets of mammals. Ellipses in the MDS space of animals (left) mark one-standard-deviation contours for the set of all mammals (thick), a representative sample (horse, chimp, seal , **A**), a somewhat representative sample (horse, mouse, rhino , **B**), and a less representative sample (horse, cow, rhino , **C**). Scatter plots (right) compare strength judgments for 45 arguments with the predictions of four models (see text).

animals. In contrast, the similarity-based models have no rational grounding and take on very different forms in the two domains. They achieve high correlations, but only through the introduction of multiple free parameters, such as the feature weights on the coin flip data, or ad hoc assumptions, such as the choice of max-similarity over sum-similarity on the animal data. On the other hand, similarity-based models do have the advantage of requiring only simple computations. Thus both Bayesian and similarity-based models may have something to offer, but at different levels of analysis. Similarity may provide a reasonable way to describe the psychological mechanisms of representativeness, while a Bayesian analysis may provide the best explanation of why those mechanisms work the way they do: why different features of sequences are weighted as they are in the coinflip example, or why max-similarity provides a better model for inductive reasoning than does sum-similarity.

Conclusion

We have argued that representativeness is best understood as a Bayesian computation, rather than as a judgment of similarity or likelihood. Our analysis makes precise one core sense of representativeness – the extent to which something is a good example of a category or process – and exposes its underlying rational basis. Rational models have been successfully applied to a number of cognitive capacities (Shepard, 1987; Anderson, 1990; Oaksford & Chater, 1998) but not previously to analyzing representativeness, which is traditionally thought of as an alternative to normative probabilistic judgment. By clarifying the relation between our intuitive sense of representativeness and normative principles of statistical inference, our analysis may lead to a better understanding of those conditions under which human reasoning may actually be rational or close to rational, as well as those situations in which it truly deviates from a rational norm.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Erlbaum, Hillsdale, NJ.

Fitelson, B. (2000). A Bayesian account of independent evidence with applications. Available at http://philosophy.wisc.edu/fitelson/psa2.pdf.

Gigerenzer, G. and Hoffrage, U. (1995). How to improve bayesian reasoning without instruction: frequency formats. *Psychological Review*, 102:684–704.

Good, I. J. (1950). *Probability and the weighing of evidence*. Charles Griffin & Co., London.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford and N. Chater, editors, *Rational models of cognition*, pages 248–274. Oxford University Press, Oxford.

Kahneman, D. and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cog. Psych.*, 3:430–454.

Kahneman, D. and Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, 103:582–591.

Mervis, C. B. and Rosch, E. (1981). Categorization of natural objects. *Annual review of psychology*, 32:89–115.

Minka, T. P. (1998). Inferring a gaussian distribution. http://www-white.media.mit.edu/ tpminka/papers/gaussian.html.

Oaksford, M. and Chater, N. (1998). *Rational models of cognition*. Oxford University Press, Oxford.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., and Shafir, E. (1990). Category-based induction. *Psychological Review*, 97:185–200.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

Rips, L. J. (1975). Inductive judgments about natural categories. *J. Verbal Learning and Verbal Behav.*, 14:665–681.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237:1317–1323.

Sloman, S. A. (1993). Feature-based induction. *Cog. Psych.*, 25:231–280.

Tversky, A. and Kahneman, D. (1983). Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90:293–315.

Acknowledgements. Supported by Mitsubishi Electric Research Labs and a Hackett studentship to TLG. N. Davidenko, M. Steyvers, and M. Strevens gave helpful comments.

A connectionist account of the emergence of the literal-metaphorical-anomalous distinction in young children

Michael S. C. Thomas (m.thomas@ich.ucl.ac.uk)
Neurocognitive Development Unit, Institute of Child Health, 30, Guilford Street
London, WC1N 1EH, UK

Denis Mareschal (d.mareschal@bbk.ac.uk)
Center for Brain and Cognitive Development, Birkbeck College, Malet Street
London, WC1E 7HX, UK

Andrew C. Hinds
Department of Psychology, King Alfred's College, Sparkford Road
Winchester, SO22 4NR, UK

Abstract

We present the first developmental computational model of metaphor comprehension, which seeks to relate the emergence of a distinction between literal and non-literal similarity in young children to the development of semantic representations. The model gradually learns to distinguish literal from metaphorical semantic juxtapositions as it acquires more knowledge about the vehicle domain. In accordance with Keil (1986), the separation of literal from metaphorical comparisons is found to depend on the maturity of the vehicle concept stored within the network. The model generates a number of explicit novel predictions.

Introduction

Despite the highly imaginative and figurative way in which children often describe the world, somewhat surprisingly it has been claimed that children are unable to understand figurative or metaphorical speech until they are quite old (Piaget, 1962; see Gibbs, 1994, for a complete review of this position). A likely explanation of this disparity is that adult usage of figurative devices such as metaphor involves several skills. For metaphor, these may include the perception of similarity and of anomaly in comprehending metaphors, the invention of similarities in generating metaphors, an understanding of the role of context in constraining possible meanings, an understanding of speaker intentions, and a metalinguistic ability to justify metaphor use based on specific cross-domain similarities (see e.g., Dowker, 1999). Moreover, it is possible that these skills have different developmental trajectories. Thus Dowker (1999) argues that age variations in similarity recognition and invention may be due to limited domain knowledge which serve to restrict the types of similarity employed by young children to mainly perceptual information. On the other hand, the temporary reduction found in the prevalence of metaphor in the language of children around the age of 6 to 7 (Gardner, Winner, Becchoffer, & Wolf, 1978; Winner, 1988) may be due to age variations in recognition of anomaly.

The idea that conceptual knowledge constrains the ability to use language figuratively is supported by evidence that metaphor usage in children is more prevalent in domains with which they are more familiar (Gottfried, 1997). Indeed Keil (1986) argued that metaphor usage closely shadows the development of conceptual categories. Similar arguments have been made in the related field of analogical reasoning, where it was also initially maintained that the relevant skills appear late in childhood (Piaget, 1962). However, when analogical reasoning was tested in more familiar domains, skills were found at a much earlier age. This implies that limitations in analogical reasoning arise from differences in the knowledge available to children as a basis for exercising this skill (Goswami, in press).

How, then, are we to interpret the apparent presence of metaphor in young children, for example, when a child aged 3 years and 5 months refers to a green carpet as 'grass' (Billow, 1981)? Putting aside the possibility of renaming in symbolic play (which need not involve any similarity between label and assigned referent), and the possibility that this is a case of over-extension (which can be ruled out by checking that the child knows the actual name for a carpet; see Gardner et al., 1978), the juxtaposition would qualify as metaphoric only under the following conditions: the child had not only spotted the similarity between the carpet and grass, but was also aware that carpet and grass fall into separate categories, so that the similarity between them was understood to be *non-literal*. Several authors have suggested that fuzziness in categorization could explain children's early use of apparently figurative language (Hakes, 1982; Marschark & Nall, 1985). If a child's conceptual knowledge has not formed into neat clusters, then there will be some overlap between categories. A sentence that appears figurative to adults may be interpreted as literal by the child.

Evidence to support this position can be found in a study by Vosniadou and Ortony (1983). In their investigations of

the emergence of the distinction between literal, metaphorical, and anomalous comparisons, these authors found evidence that, although 3-year-olds could produce metaphorical completions to target sentences, they were unable to reliably identify that the concepts juxtaposed in these sentences fell into separate categories. However, by four years of age, children who produced metaphors also showed an understanding that metaphorical statements involved concepts from different conventional categories. Both the 3- and 4-year-olds were able to identify anomalous from literal and metaphorical comparisons (see also Pearson, 1990). Vosniadou and Ortony interpreted these data as suggesting that children start with an undifferentiated notion of similarity, which at about the age of four becomes differentiated into literal and non-literal similarity. They suggested that the latter type forms the basis of metaphorical language comprehension.

In this paper, we describe the first computational model explaining the emergence of the distinction between literal and metaphorical similarity, based on an existing connectionist model of simple metaphor comprehension (Thomas & Mareschal, 2001). The importance of this model is that it directly relates the development of metaphor comprehension to the development of semantic representations. The structure of this paper is as follows. We begin by briefly reviewing connectionist approaches to metaphor comprehension. Second, we describe the main tenets of the Metaphor by Pattern Completion (MPC) model on which the developmental account is based. Third, we chart the development of category-specific representations that support metaphor comprehension and the distinction between literal and figurative statements within the MPC model. Finally, we discuss implications for the order of acquisition of such distinctions by young children.

Connectionist models of metaphor processing

First of all, it is important to point out that, although previous computational models have been proposed for the comprehension of metaphor, all of these models have related to the adult state, and none have contained a developmental component.

Previous models of metaphor comprehension have exploited the soft multiple constraint satisfaction abilities of connectionist networks to capture the interactions of conceptual domains when they are juxtaposed in comparisons. One class of models has focused on the potential of microfeature or vector representations of concepts to capture subtle interactions between knowledge bases (e.g., Chandler, 1991; Sun, 1995; Thomas & Mareschal, 2001). A second class of models has focused on structural mapping accounts of analogy formation, whereby target and vehicle domains are compared via the alignment of their relational structure, as well as evaluation of shared attributes (e.g., Holyoak & Thagard, 1989; Hummel & Holyoak, 1997). Why have computational models of metaphor comprehension been silent on developmental phenomena? The answer is that both classes of model have tended to include extensively pre-structured, domain-specific representations, which prevent them from exploring how representations (and their comparison) may emerge as a function of development.

In the present work, we will focus only on attribute mapping, which is readily captured by microfeature models, and put to one side problems of structural alignment. Although this limits the scope of the metaphors to which the model can be applied, it nevertheless makes the first initial steps towards exploring the developmental dimension of metaphor processing, and specifically, to investigating the ways in which metaphor comprehension can be linked to the development of semantic representations.

The MPC model

A full description of the MPC model can be found in Thomas and Mareschal (2001), along with an evaluation of its main assumptions. Here we provide a brief outline. In broad terms, the model suggests that, when presented with a metaphor such as *Richard is a lion*, the listener indeed attempts to fit the concept *Richard* into the category of *lion*; in so doing, an outcome of the categorization process is to alter the representation of *Richard* to make him more consistent with the features of a lion.

More specifically, metaphor comprehension is construed as a two-stage process. Consistent with Glucksberg and Keysar's (1990) view of metaphor comprehension as a type of categorization process, the first stage comprises *mis-classification* of a semantic input. A metaphor <A is B>, where A is the *topic* and B the *vehicle*, is comprehended by applying a representation of the first term (A) to a semantic memory network storing knowledge about the second term (B). Categorization is evaluated via the accuracy of reproduction of (A)'s representation in an autoassociator network trained on exemplars of (B). The degree of semantic distortion of (A) is a measure of the semantic similarity of concept A to domain B (Thomas & Mareschal, 2001).

However, the result of applying (A) to the network storing knowledge about (B) is a representation of (A) transformed to make it more consistent with the (B) knowledge base. In particular, there is an interaction in which features of (A) key into covariant structure between features in (B). If (A) shares some features of such covariant structures, it inherits further features by a process of *pattern completion*. Such feature inheritance depends on both terms, and provides an implementation of Black's (1979) well-known interaction theory of metaphor comprehension. However, enhancement of the features of (A) does not complete the process. In a second stage, the degree of meaning change of the topic is compared to the expected level of change given the current discourse context (Vosniadou, 1989). If the threshold is high, the statement is taken as literal and the full change in meaning is accepted. If it is at an intermediate level, only enhanced feature changes are accepted as the

communicative intent of a metaphor. If the threshold is at a low level, the sentence is rejected as anomalous.

Thomas & Mareschal (2001) evaluated the model's performance in comparing highly simplified domains to illustrate this process. Plausible metaphorical comparisons such as "the apple is a ball" were contrasted with anomalous comparisons such as "the apple is a fork". The model was able to account for a number of empirical phenomena, including the non-reversibility of comparisons and the predictability of interactions between topic and vehicle.

However, the degree to which metaphorical semantic transformations will occur depends not only on the similarity of (A) and (B), but also on the amount and quality of the knowledge stored in knowledge base B. In this way, metaphor comprehension can be linked to semantic development.

In the next section, we take a single vehicle knowledge domain and trace the development of metaphorical comprehension as the knowledge in the base network increases with learning. For simplicity, the sample knowledge base comprises information about *types of ball*, and performance is compared on literal comparisons ("the football is a ball") against metaphorical comparisons ("the pumpkin is a ball") and anomalous comparisons ("the kite is a ball").

The developmental model is intended to be illustrative: we make no claims about children's specific abilities to compare objects to balls at specific ages. Rather, we are interested in evaluating the effect of emerging semantic structure on the delineation of different types of similarity, and the consequent qualitative changes in the nature of metaphor comprehension during development.

Modeling the development of metaphor comprehension

Autoassociation is at the heart of the MPC mechanism. In the original model (Thomas & Mareschal, 2001), multiple parallel knowledge bases were available for different comparisons. However, in the present article and in the interest of clarity, we discuss only results obtained with a single autoassociator network.

A network with 16 input units, 16 output units, and 10 hidden units was trained to autoassociate a set of input patterns that defined the semantic knowledge of the vehicle domain. The number of hidden units was chosen to allow good training performance but also to encourage generalization. All units in the network used sigmoid activation functions.

The autoassociation network was trained for 500 presentations of the complete training set. At each epoch the training set was presented in a different random order. The learning rate and momentum were set to 0.05 and 0.0 respectively. Metaphor comprehension performance was evaluated at 0, 1, 2, 3, 4, 5, 7, 10, 15, 20, 30, 45, 70, 110, 200, and 500 epochs of training. The results reflect an average over n=12 replications with different initial random seeds.

The training set was constructed around 8 prototypes of various balls, constituting the 'ball' knowledge base. Prototypes were defined over 5 clusters of features: color (red, green brown, white), shape (round, irregular), consistency (soft, hard), size (small, large), weight (heavy, light), and associated action (thrown, kicked, hit, eaten), for a total of 16 semantic features. The last feature was included to permit anomalous and metaphorical comparisons. We assume that all concepts can be described by the same large feature set, and that the organization of knowledge into different categories happens within the hidden unit representations through learning. Feature values ranged between 0 and 1, so that the higher the activation, the more prominent the feature. Opposite feature values (e.g., small & large) were encoded on separate inputs to allow the coding of an absence of knowledge. From each prototype, 10 exemplars were generated by adding Gaussian noise (with standard deviation of 0.35) to the prototype pattern. The final training set thus constituted 8x10 = 80 exemplars of balls. The training prototypes are listed in Table 1, upper section.

Assessing different semantic comparisons

A comparison is evaluated by applying a novel input to the network and seeing how well it is reproduced on the output units. The more accurate the reproduction, the greater the similarity of the novel item to the knowledge stored within the network. Nine novel comparisons were created using the semantic features described above. These fell into three classes: (1) literal comparisons, (2) metaphorical comparisons, and (3) anomalous comparisons.

Literal comparisons involved novel exemplars of balls near the prototypical values. Metaphorical comparisons involved inputs that shared some properties with balls, but differed on other properties. Anomalous comparisons involved inputs constructed so that the inputs shared few features with balls in general.

The input vectors for the different classes of comparisons were constructed by comparing the novel input with the ball prototype vectors used to generate the knowledge training set. This was achieved by computing the angle between the two vectors in semantic space and selecting the closest match. For the literal comparisons, the angle had to be to be less than 10 degrees, for the metaphorical comparisons, it had to be between 40 and 45 degrees, and for the anomalous comparisons, it had to be between 60 and 66 degrees. (An angle of 90 degrees would constitute a novel pattern orthogonal to, or completely different from, all the prototypes used to generate the exemplars in the knowledge base.) Novel comparisons are shown in Table 1, lower section. A perfect reproduction of the input at the output indicates a similarity of 1.0 (self-similarity). The transformation similarity (*S*) of each novel comparison to the ball knowledge base was defined as:

$$S = 1 - RMS\ Error \quad (1)$$

An RMS error of 0 would give a similarity of S=1. High similarity implies low semantic distortion (as expected in a literal comparison), moderate similarity implies moderate semantic distortion (as expected in a metaphorical comparison), and low similarity implies high semantic distortion (as expected in an anomalous comparison). The similarity of novel comparisons was evaluated at different points during training. Principal Component Analyses of the hidden unit activations were also carried out during training to chart the development of the internal representations.

Results

Figure 1 shows S for each of the three types of comparison as learning progresses. Initially, there is little difference between literal, metaphorical and anomalous comparisons. However, even very early in learning a marked separation of the anomalous comparisons from the literal and metaphorical comparisons appears. The metaphorical and literal comparisons continue to be treated in a similar fashion for a further 5 presentations of the training set. At this point, metaphorical and literal similarities diverge. In the remaining epochs of training, the similarities from the three different types of comparisons separate into distinct bands. After an initial period of treating literal and metaphorical statements identically, the network has learnt to separate them out.

The process that underlies the development of this distinction can be better understood by examining the developing structure of the network knowledge base (Fig.2). Principal Components Analysis of the hidden unit activation space shows how the internal representations pull apart the different types of ball during training, according to their input characteristics.

In general, anomalous patterns fall in-between clusters, while metaphorical comparisons lie at the edge of clusters, and literal comparisons lie within the clusters. Once the clusters are sufficiently delineated from each other, an item that bears a metaphorical relation to a given category is distinguished from members of that category.

Novel inputs to the network are transformed in an attempt to classify them. Within the model, the transformed semantic representation corresponds to the meaning enhancement that is the outcome the comparison. Focusing on the metaphorical comparisons alone, examination of this enhancement yields three distinct phases during training. First, there is poor pattern completion, linked to an immature vehicle knowledge base. Next, with the initial emergence of semantic structure, metaphorical comparisons such as “the pumpkin is a ball” and “the apple is a ball” lead to enhancement of some of the target’s features. For example, ‘pumpkins’ and ‘apples’ are not associated with being ‘thrown’, ‘hit’, or ‘kicked’. The effect of each metaphor is to transfer such features from vehicle to topic. However, initially enhancement occurs according to an early, prototypical notion of ball, a notion that averages over all exemplars of balls, and corresponds to what one might call the *basic level* of the category. On average, most balls are ‘hit’ rather than ‘kicked’ or ‘thrown’. During this second phase, the ‘hit’ enhancement is inherited by all round, firm targets such as ‘apple’ and ‘pumpkin’. However, in the third phase, as further training produces delineation of the knowledge base, transfer now occurs according to the type of ball *most similar* to the particular target, according to

Table 1: Upper section: Prototypical patterns forming the ball knowledge base. Adding noise to the prototypes creates training sets. Lower section: Novel patterns used in literal, metaphorical, and anomalous comparisons.

	Color				Action				Shape		Consist-ency		Size		Weight	
Prototypes	Red	Green	Brown	White	Eaten	Thrown	Hit	Kicked	Round	Irregular	Soft	Hard	Small	Large	Heavy	Light
Football(white)	.00	.00	.00	.90	.00	.20	.00	.95	.90	.00	.00	.80	.00	.90	.90	.00
Football(brown)	.00	.00	.90	.00	.00	.20	.00	.95	.90	.00	.00	.80	.00	.90	.90	.00
Cricket ball	.90	.00	.00	.00	.00	.98	.97	.00	.90	.00	.00	.98	.80	.00	.80	.00
Ping-Pong ball	.00	.00	.00	.95	.00	.10	.98	.00	.98	.00	.00	.98	.95	.00	.00	.95
Tennis ball	.00	90	.00	.00	.00	.80	.98	.00	.90	.00	.80	.00	.80	.00	.00	.85
Squash ball (red)	.80	.00	.00	.00	.00	.50	.98	.00	.93	.00	.85	.00	.95	.00	.00	.90
Squash ball (green)	.00	.90	.00	.00	.00	.50	.98	.00	.93	.00	.85	.00	.95	.00	.00	.90
Beach ball	.98	.00	.00	.00	.00	.90	.90	.90	.90	.00	.98	.00	.00	.98	.00	.90
Novel comparisons																
Literal:																
Football (white)	.00	.00	.00	.85	.00	.20	.00	.98	.80	.00	.10	.80	.00	.90	.80	.00
Beach ball	.90	.00	.00	.00	.00	.80	.70	.90	.70	.00	.90	.00	.00	1.0	.00	.80
Ping-Pong ball	.00	.00	.00	.99	.00	.20	.99	.00	.95	.00	.00	.90	.95	.00	.00	.97
Metaphorical:																
Apple (red)	.80	.00	.00	.00	.95	.05	.00	.00	.75	.15	.70	.20	.70	.00	.00	.50
Pumpkin	.20	.00	.70	.00	.80	.00	.00	.00	.80	.50	.80	.60	.00	.80	.90	.00
Apple (green)	.00	.95	.00	.00	.95	.05	.00	.00	.75	.15	.70	.20	.70	.00	.00	.50
Anomalous:																
Kite	.99	.00	.00	.00	.00	.05	.00	.00	.00	.99	.00	.98	.00	.95	.20	.80
Spaghetti	.00	.00	.80	.20	.97	.00	.00	.00	.00	.70	.80	.20	.00	.70	.00	.60
Toast	.00	.00	.80	.10	.80	.00	.00	.00	.00	.80	.80	.00	.80	.00	.00	.90

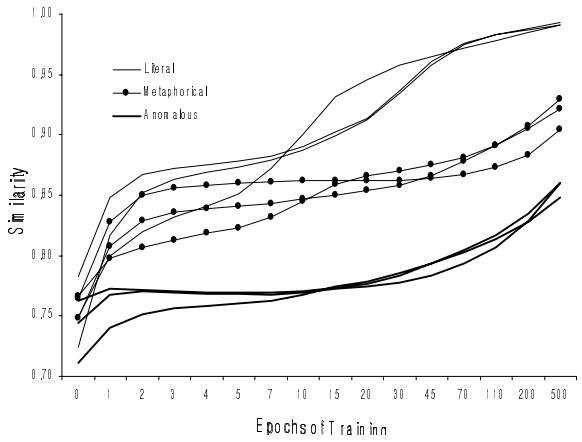


Figure 1. Similarity (*S*) of novel comparisons to the ball knowledge base during training. Three examples from each comparison type are plotted.

what one might call the *subordinate level* of the vehicle category. Table 2 shows that at 4 epochs 'apple' and 'pumpkin' have similar activation levels for the action features, loading maximally on 'hit', whereas at 500 epochs, 'apple' and 'pumpkin' now load on different features. Apples are now viewed as likely to be hit, and pumpkins to be kicked, according to their differing sizes. The model thus generates an explicit and testable prediction: *attribute inheritance will move from basic to subordinate level during development*.

Moreover, since there is variability within the internal structure of categories, not all literal comparisons will be equivalent. The more atypical the literal comparison, the more it will resemble a metaphor. This leads to a second explicit and testable prediction: *the recognition of atypical literal statements as distinct from metaphorical statements should lag behind the recognition of typical literal statements as distinct from metaphorical statements during development.*

Discussion

A common characterization of conceptual development views young children's knowledge as being assimilated into broad groups; as children develop, they make finer and finer distinctions until there are many different categories (e.g., Carey, 1985; Keil, 1986). Because the comprehension of metaphor requires the deliberate deconstruction of

Table 2: Attribute transfer from *basic* (4 epochs) and *subordinate* (500 epochs) category levels. Scores show the transformed feature values for action features Thrown (T), Hit (H) and Kicked (K) in the topic.

	Comparison <X is a Ball>					
	Red Apple			Pumpkin		
	T	H	K	T	H	K
4 epochs	.59	.75	.37	.50	.60	.30
500 epochs	.17	**.85**	.17	.18	.03	**.48**

a) 4 epochs

b) 15 epochs

c) 500 epochs

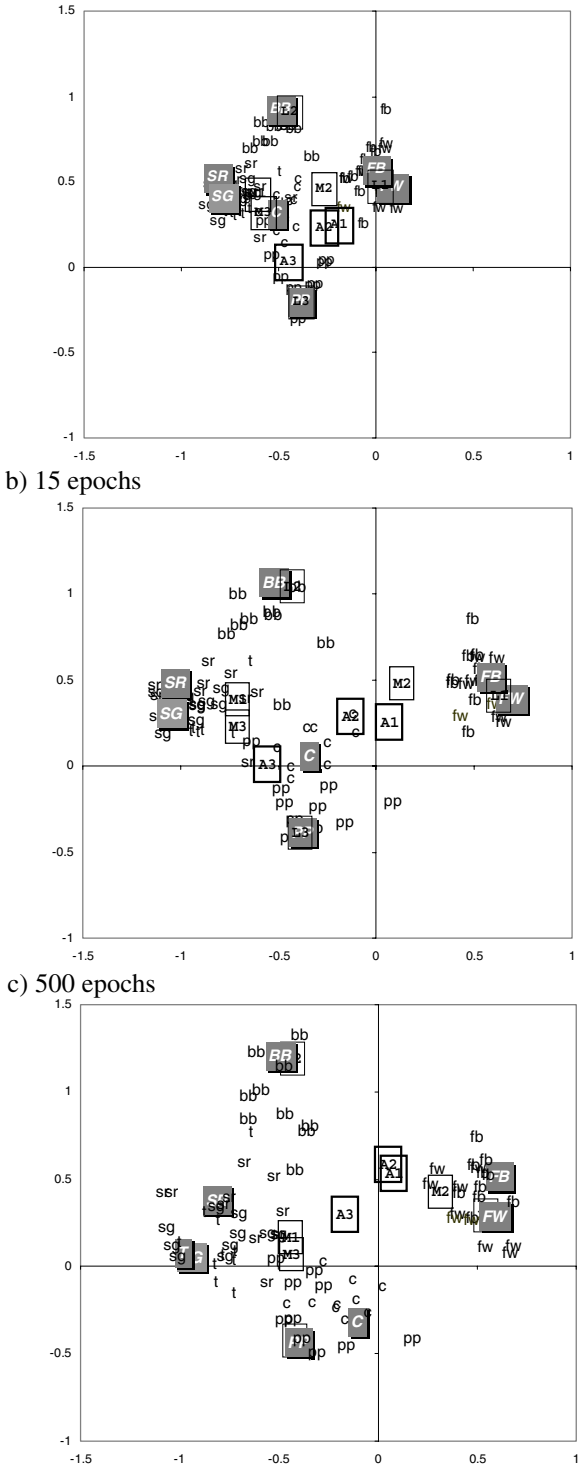


Figure 2. First two components of the hidden unit activations for training and test patterns of a representative network across training.

categories, the way knowledge is categorized will have a large effect on metaphor comprehension. The model we have described above provides a concrete implementation of Marschark and Nall's (1985) account of metaphor use in young children. Literal, metaphorical, and anomalous comparisons fall onto a conceptual space undergoing refinement. The process of refinement leads to the emergence of a notion of non-literal similarity.

Clearly this simple model does not capture all aspects of the development of metaphor comprehension. The metaphors we have dealt with are predominantly perceptual. Importantly, the model fails to capture the emerging use of structural information in children's metaphors (Gentner, 1988). However, existing computational models have not addressed develop-mental phenomena at all, let alone the relational shift. The next step for the MPC model will be an extension to structured representations, possibly via the inclusion of synchrony binding (see Hummel & Holyoak, 1997), while retaining the mechanism of pattern completion as a powerful tool for explaining the transfer of attributes in metaphorical comparisons. Despite its simplicity, the importance of the current model is its demonstration that the emergence of non-literal similarity can be driven by emerging semantic structure, and the explicit testable hypotheses it generates to progress our understanding of the development of metaphor comprehension in young children.

Acknowledgements

MT's work on this paper was funded by MRC Project Grant no. G9809880 and MRC Programme Grant no. G9715642; DM's work was funded in part by European Commission RTN Grant CT-2000-00065 and ESRC Grant no. R000239112.

References

Billow, RM (1981). Observing spontaneous metaphor in children. *Journal of Experimental Child Psychology*, *31*, 430-445.

Black, M (1979). More about metaphor. In A. Ortony (Ed.). *Metaphor and thought*. Cambridge, UK: Cambridge University Press.

Carey, S (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Chandler, SR (1991). Metaphor comprehension: A connectionist approach to implications for the mental lexicon. *Metaphor & Symbolic Activity*, *6*, 227-258.

Dowker, A (1999). Metaphor: Is it the same for children and adults? In *Proceedings of the AISB'99 Symposium on Metaphor, Artificial Intelligence, and Cognition.* AISB: Brighton, UK.

Gardner, H, Winner, E, Becchoffer, R & Wolf, D (1978). The development of figurative language. In K. Nelson (Ed.) *Children's language Vol.1*. Cambridge, MA: Gardner Press Inc.

Gentner, D (1988). Metaphor as structure mapping: The relational shift. *Child Development*, *59*, 47-59.

Gibbs, RW (1994). *The poetics of mind.* Cambridge, UK: Cambridge University Press.

Glucksberg, S & Keysar, B (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review*, *97*, 3-18.

Goswami, U (in press). Analogical reasoning in children. In D. Gentner, K.J. Holyoak, Kokinov (Eds.). Analogy: Interdisciplinary perspectives.

Gottfried, GM (1997). Comprehending compounds: evidence for metaphoric skill? *Journal of Child Language*, *24*, 163-186.

Hakes, D (1982). The development of metalinguistic awareness: What develops? In S. Kuczaj (Ed.) *Language development Vol.2*. Hillsdale N.J.: Erlbaum

Holyoak, KJ & Thagard, P (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, *13*, 295-355.

Hummel, JE & Holyoak, KJ (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*, 427-466.

Keil, FC (1983). On the emergence of semantic and conceptual distinctions. *Journal of Experimental Psychology: General*, *112*, 357-385.

Marschark, M & Nall, C (1985). Metaphor competence in cognitive and language development. *Advances in Child Development and Behavior*, *26*, 49-78.

Pearson, B (1990). The comprehension of metaphor by preschool children. *Journal of Child Language*, *17*, 185-203.

Piaget, J (1962). *Play, dreams, and imitation in childhood*. New York, NY: Norton.

Sun, R (1995). A microfeature based approach towards metaphor interpretation. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (pp. 424-429).

Thomas, MSC & Mareschal, D (2001). A connectionist model of metaphor by pattern completion. *Metaphor & Symbol*, *16*, 5-27.

Vosniadou, S (1989). Context and the development of metaphor comprehension. *Metaphor & Symbolic Activity*, *4*, 159-171.

Vosniadou, S & Ortony, A (1983). The emergence of the Literal-Metaphorical-Anomalous distinction in young children. *Child Development*, *54*, 154-161.

Winner, E (1988). *Invented worlds: The psychology of the arts.* Cambridge, MA: Harvard University Press.

A New Model of Graph and Visualization Usage

J. Gregory Trafton
Naval Research Laboratory
trafton@itd.nrl.navy.mil

Susan B. Trickett
George Mason University
stricket@gmu.edu

Abstract

We propose that current models of graph comprehension do not adequately capture how people use graphs and complex visualizations. To investigate this hypothesis, we examined 3 sessions of scientists using an *in vivo* methodology. We found that in order to obtain information from their graphs, scientists not only read off information directly from their visualizations (as current theories predict), but they also used a great deal of mental imagery (which we call spatial transformations). We propose an extension to the current model of visualization comprehension and usage to account for this data.

Introduction

If a person looks at a standard stock market graph or a meteorologist is examining a complex meteorological visualization, how is information extracted from these graphs? The most influential research on graph and visualization comprehension is Bertin's (1983) task analysis that suggests three main processes in graph and visualization comprehension:

1. Encode visual elements of the display: For example, identify lines and axes. This stage is influenced by pre-attentive processes and is affected by the discriminability of shapes.

2. Translate the elements into patterns: For example, notice that one bar is taller than another or the slope of a line. This stage is affected by distortions of perception and limitations of working memory.

3. Map the patterns to the labels to interpret the specific relationships communicated by the graph. For example, determine the value of a bar graph.

Most of the work done on graph comprehension has examined the encoding, perception, and representation of graphs. Cleveland and McGill, for example, have examined the psychophysical aspects of graphical perception (Cleveland & McGill, 1984, 1986). Similarly, Pinker's theory of graph comprehension, while quite broad, focuses on the encoding and understanding of graphs (Pinker, 1990). Kosslyn's work emphasizes the cognitive processes that make a graph more or less difficult to read. Kosslyn's syntactic and semantic (and to a lesser degree pragmatic) level of analysis focuses on encoding, perception, and representation of graphs (Kosslyn, 1989). Recent work by Carpenter and Shah (1998) shows that people switch between looking at the graph and the axes in order to comprehend the visualization.

This scheme seems to work very well when the graph contains all the information the user needs (i.e., when the information is explicitly represented in one form or another). Thus, when an undergraduate is asked to extract specific information from a bar-graph, the above process seems to hold. However, graph usage outside the laboratory is probably not simply a series of information extractions. For example, when looking at a stock market graph, the goal may not be just to determine the current or past price of the stock, but perhaps to determine what the price of the stock will be sometime in the future. A weather forecaster looking at a meteorological visualization is frequently trying to predict what the weather will be in the future, as well as what the current visualization shows (Trafton, Kirschenbaum, Tsui, Miyamoto, Ballas, & Raymond, 2000). A scientist examining results from a recent experiment can not always display the available information in a way that perfectly shows the answer to her hypotheses.

How do current theories of graph comprehension hold up when a graph or visualization does not contain the exact information needed? Unfortunately, the theories do not say anything about this situation. In fact, there are no specifications in any theory of graph comprehension about how information could or would be extracted from a visualization where that information is not represented in some form. If a graph does not contain the information needed by the user, the graph is often labeled "bad" or "useless" (Kosslyn, 1989; Pinker, 1990).

Current graph comprehension theories do not have a great deal to say about what to do when a graph does not explicitly show the needed information for a variety of reasons. The main reason is probably that most graph comprehension studies have used fairly simple graphs for which no particular domain knowledge is required (e.g., Carter, 1947; Lohse, 1993; Pinker, 1990). However, in real-world situations, people use complex visualizations that require a great deal of domain knowledge, and all the needed information would probably not be explicitly represented in the graph. This study will thus try to answer two questions about graph comprehension. Do expert users of visualizations ever need information that is not on a specific graph they are using? If so, how do they extract that information from the graph?

There are several possible things that users could do when trying to extract information from a graph. In the simplest case, the information is explicitly available, and they can simply read off the information from the visualization. What do they do when information they need is not available on the visualization? They could create a completely new visualization that does show the information. They could also collect more data or consult another source. They could create an explicit plan to look for more data or run another experiment.

What do they do when the visualization is all they have to work with? What kind of mental operations could users perform on graphs and visualizations in order to extract information that is not explicit? One possibility is that people use some sort of visual imagery to extract information that is not explicitly represented on a graph or visualization. For example, a weather forecaster may mentally imagine a front moving east over the next several days (Trafton et al., 2000), or a stock analyst may mentally extend a line on a graph and think that a stock will continue to rise. We have developed a framework for coding and working with these kinds of graphs and visualizations called *Spatial Transformations* that will be used to investigate these issues. We will argue that spatial transformations are a fundamental aspect of complex visualization usage.

Spatial Transformations are cognitive operations that a scientist performs on a visualization. Sample spatial transformations are mental rotation (e.g., Shepard & Metzler, 1971), creating a mental image, modifying that mental image by adding or deleting features to or from it, animating an aspect of a visualization (Hegarty, 1992) time series progression prediction, mentally moving an object, mentally transforming a 2D view into a 3D view (or vice versa), comparisons between different views (Kosslyn, Sukel, & Bly, 1999; Trafton, Trickett, & Mintz, 2001), and anything else a scientist mentally does to a visualization in order to understand it or facilitate problem solving. Also note that a spatial transformation can be done on either an internal (i.e., mental) image or an external image (i.e., a scientific visualization on a computer-generated image). What all spatial transformations have in common is that they involve the use of mental imagery. A more complete description of spatial transformations can be found at `http://iota.gmu.edu/users/trafton/405st.html`.

We will examine the number of times that users needed information from a visualization. If all or most of the information is available explicitly on the visualization, we should see primarily read-offs (Kosslyn, 1989; Pinker, 1990). If, however, a particular visualization does not explicitly display particular information that a scientist wants, we will examine how the scientist goes about obtaining that information. We expect that in complex visualizations, there is a great deal of information that is needed in addition to what is displayed, and we expect scientists to use spatial transformations to retrieve that information.

Method

In order to investigate the issues discussed above, we have adapted Dunbar's in vivo methodology (Dunbar, 1995, 1996; Trickett, Trafton, & Schunn, 2000b). This approach offers several advantages. First, it allows the observation of experts, who are thus able to use their domain knowledge to guide their strategy selection. Second, it allows the collection of "on-line" measures of thinking, which allow the investigation of the scientists' reasoning as it occurs (Ericsson & Simon, 1993). Finally, the tasks (experiment design, data analysis, etc.) conducted by the scientists, as well as the tools they use, are fully authentic.

Two sets of scientists were videotaped while conducting their own research. All the scientists were experts, having earned their Ph.D.s more than 6 years previously. In the first set, two astronomers, one a tenured professor at a university, the other a fellow at a research institute, worked collaboratively to investigate computer-generated visual representations of a new set of observational data. At the time of this study, one astronomer had approximately 20 publications in this general area, and the other approximately 10. The astronomers have been collaborating for some years, although they do not frequently work at the same computer screen and the same time to examine data.

In the second dataset, a physicist with expertise in computational fluid dynamics worked alone to inspect the results of a computational model he had built and run. Two related sessions were recorded with this scientist over consecutive days. He works as a research scientists at a major U.S. scientific research facility, and had earned his Ph.D. over 20 years previously. He had inspected the data previously but had made some adjustments to the physics parameters underlying the model and was therefore revisiting the data.

Both sets of scientists were instructed to carry out their work as though no camera were present and without explanation to the experimenter (Ericsson & Simon, 1993). The relevant part of the astronomy session lasted about 53 minutes, and the two physics sessions each lasted approximately 15 minutes. All utterances were later transcribed and segmented according to complete thought. All segments were coded by 2 coders as on-task (pertaining to data analysis) or off-task (e.g., jokes, phone call interruptions, etc.). Inter-rater reliability for this coding was more than 95%. Off-task segments were excluded from further analysis. On-task segments (N = 649 for the astronomy dataset and N = 189 for the first physics dataset and N = 176 for the second physics dataset) were further coded as described below.

The Tasks and the Data

Astronomy The astronomical data under analysis were optical and radio data of a ring galaxy. The astronomers' high-level goal was to understand its evolution and structure by understanding the flow of gas in the galaxy. In order to understand the flow of gas, the astronomers must make inferences about the velocity field, represented by

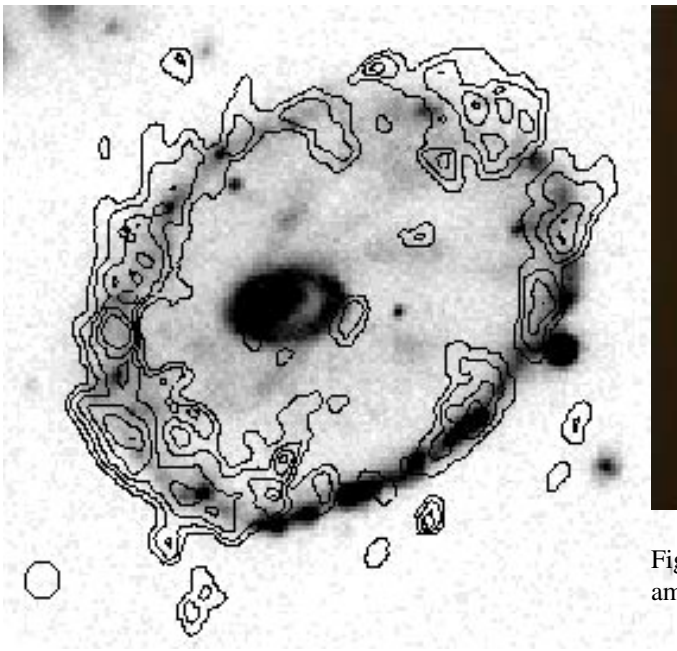


Figure 1: An example of the kind of visualizations examined by the astronomers.

contour lines on the 2-dimensional display. The astronomers' task was made difficult by two characteristics of their data. First, the data were one- or at best 2-dimensional, whereas the structure they were attempting to understand is 3-dimensional. Second, the data were noisy, and there was no easy way to distinguish between noise and real phenomena. Figure 1 shows a screen snapshot of the type of data the astronomers were examining. In order to make their inferences, the astronomers used different types of image, representing different phenomena (e.g., different forms of gas), which represent different information about the structure and dynamics of the galaxy. Some of these images could be overlaid on each other. In addition, the astronomers could choose from images created by different processing algorithms, each with advantages and disadvantages (e.g., more or less resolution). Finally, they could adjust different features of the display, such as contrast or false color. A more complete description of this dataset can be found in Trickett, Fu, Schunn, and Trafton (2000a) and Trickett, Trafton, and Schunn (2000b).

Physics The physicist was working to evaluate how deep into a pellet a laser light will go before being reflected. His high-level goal was to understand the fundamental physics underlying the reaction, an understanding that hinged on an understanding of the relative importance and growth rates of different modes. The physicist had built a model of the reaction; other scientists had independently conducted experiments in which lasers were fired at pellets and the reactions recorded. A close match between model and empirical data would indicate a good understanding of the underlying theory. Although the physicist had been in conversation with the experimentalist, he had not viewed the empirical data, and in this session he was investigating only the results of his computational model. However, he believed the model to be correct (i.e., he had strong expectations about what he would see), and in this sense, this session may be considered confirmatory.

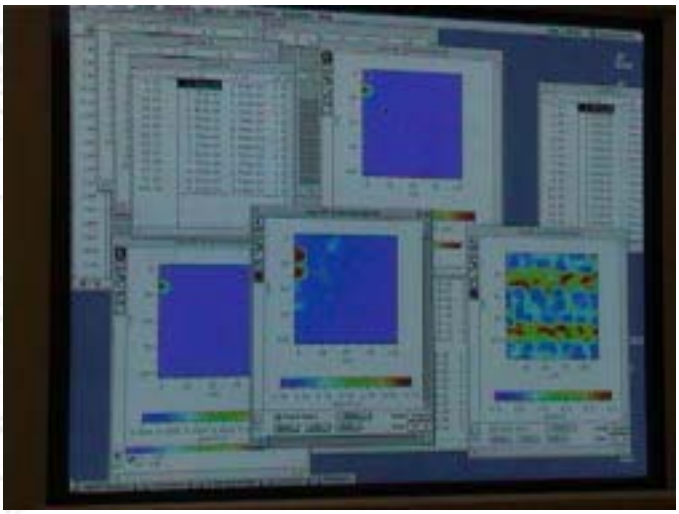


Figure 2: An example of the kind of visualizations examined by the physicist.

The data consisted of two different kinds of representation of the different modes, shown over time (nanoseconds). The physicist was able to view either a Fourier decomposition of the modes or a representation of the "raw" data.

Figure 2 shows an example of the physicist's data. He could choose from black-and-white or a variety of color representations, and could adjust the scales of the displayed image, as well as some other features. He was able to open numerous views simultaneously.

Coding Scheme

Our goals in this research are first, to determine if complex visualizations contain all the information needed by the scientists, and, if not, to investigate what happens when they do not have all the information they need. We propose that spatial transformations are a major portion of extracting information from a visualization when the data is not explicitly represented. Consequently, we identified every situation where a scientist wanted to extract information from a visualization. Next, we coded what the scientist did to extract information, including reading off the information directly from the graph, spatial transformations, changing the visualization, plans or discussions about getting more data, and abandoning their attempt to get the information. We now describe and provide examples of this coding scheme in detail.

Example	Explanation
After all, it is ten to the minus six...	Scientist is looking at a line and extracting the y-axis value
I mean, the fact you see such a strong concentration of gas in the ring, um...	Scientist is reading off the amount of gas in the ring
That's about 220 km/sec, which is the velocity spread of a normal galaxy.	Scientist is reading off the velocity spread

Table 1: Examples of information that is read off the visualizations.

Spatial Transformation	Example	Explanation
Create Mental Image	I mean, in a perfect, in a perfect world, in a perfect sort of spider diagram...	Scientist is creating a mental image of a spider diagram; there is no spider diagram displayed.
Modify Image	So that [line] would be below the black line	Scientist is adding a new (hypothesizied) line to a current visualization
Modify Image	If there was no streaming motion or sort of piling of gas	Scientist has imaged a previous mental image and is now removing the streaming motions from his mental image
Comparison:	Maybe it's a projection effect, although if that's true, there should be a very large velocity dispersion.	Scientist is comparing a current image to a previously created mental image.

Table 2: Examples of spatial transformations.

Desire to extract information A scientist would frequently want to extract some amount of information from a visualization. Comments varied from the very general ("What do we see?") to the very specific ("Let's see, how does oh-three versus three-oh [look]?").

Read-Off A scientist would be able to read-off information directly from the graph. Information that was read off a visualization was explicitly on the graph and the scientist simply had to read-off a particular value. For every utterance, we evaluated whether a value was read off the visualization. Table 1 shows several examples of information that was read off of the visualization.

Spatial Transformations As discussed earlier, spatial transformations are cognitive operations that a scientist performs on a visualization. For every utterance in each protocol we evaluated whether there was a spatial transformation. Spatial transformations were further coded as Create Image, Modify Image, or Comparison. Table 1 shows examples of each category of spatial transformation (note that these utterances are independent of one another and do not represent a sequence). Table 2 shows several examples of spatial transformations that were used by the scientists.

Changing the Visualization The scientists were using their own tools and were able to change the visualization to a completely different representation. For example, a scientist could change the data display from the raw data to a Fourier mode display). Alternatively, the scientists could "tweak" the current representation (from black and white to color, for example). We coded the visualization changes where the scientists were looking for additional information. If they simply made a mistake and tweaked the visualization, we did not count that visualization change. For example, while looking at a particularly compressed visualization, one of the scientists said "Where's three-oh at? Don't see three [oh]. That's what I figured, I was gonna get spaghetti. Let's do a re-plot." and then replotted the data with a reduced dataset.

Plans to gather more data Occasionally, the scientists wanted or needed to gather more data. We coded every time they made a plan to gather more data. For example, one scientist said "So that means that this guy is in fact between him and him, which is exactly what the experimentalist believes he saw. Now, somewhere along the line I have to get their results."

Abandoning their attempt to get information Sometimes the scientists either could not decide what data to get or simply abandoned their quest for a specific information. We coded every time the scientists abandoned their attempt to get information. For example, one scientist, unable to explain a particular feature after extensive investigation of the image, said "Yeah well, [let's] gloss over it."

Results

Our two goals in this paper are to explore whether scientists are able to directly extract the amount of information they need from the visualizations they examine and if not, to explore how they do get the information that is needed.

How often is needed information directly available?

Of the 1014 total utterances in the three sessions, almost half (481) involved some form of information gathering.

As Figure 3 shows, approximately half of those information gathering instances were read-off, suggesting that the scientist did use the visualization a great deal to extract information. However, there were many times when the scientists needed information from a visualization but it was not available directly from the visualization. Thus, the visualizations seem to be good, but far from perfect from an information gathering point of view.

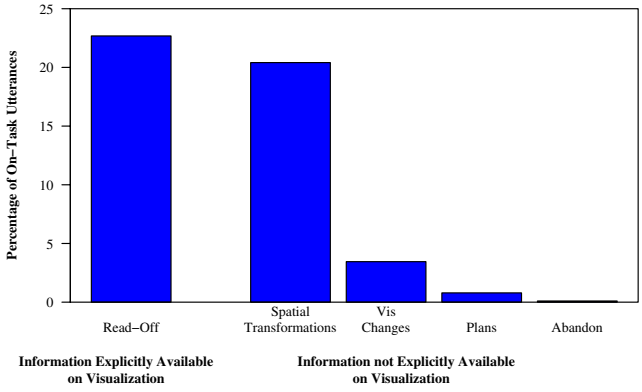


Figure 3: The number of read-offs, spatial transformations, visualization changes, plans to collect future data, and decisions to abandon the attempt to get more data for all datasets.

How was needed information extracted if it was not simply read off? As Figure 3 shows, the vast majority of information that was not read off was gathered by using spatial transformations. In fact, there was no statistical difference between the number of times that the scientists read off information directly from the graph and the number of spatial transformations, $\chi^2(1) = 1.21, p > .20$.

Additionally, scientists chose to use a spatial transformation to get needed information from a visualization rather than changing the visualization, $\chi^2(1) = 122.25, p < .001$, making plans to gather more data $\chi^2(1) = 184.19, p < .001$, or abandoning their attempt to answer their question, $\chi^2(1) = 204.02, p < .001$.[1]

[1] All χ^2's used the Bonferroni adjustment.

General Discussion

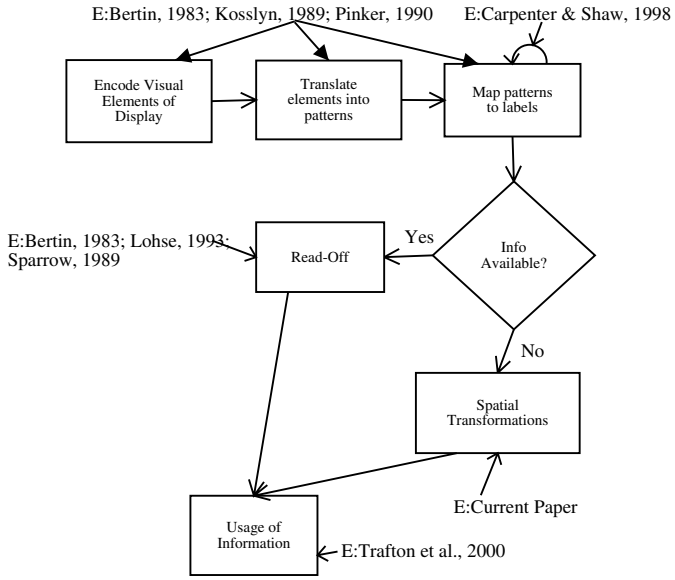


Figure 4: Our current theoretical model of complex visualization usage. The "E:" shows evidence for each stage of the model.

We have conducted a detailed analysis of expert scientists at work in their own laboratories, analyzing data that they have collected themselves. Our results show that these scientists do extract a great deal of information from the visualizations. However, these visualizations do not provide the scientists with all the information they need to answer their questions. We found that when they needed information that was not explicitly provided by the visualization, they tended to perform spatial transformations to answer their questions.

It is interesting that the scientists did not simply change the visualization more frequently to get the needed information. There was some evidence in the protocols that it was not easy to create new visualizations. For example, some of the visualizations had to be re-done because of an error that was made in the display (i.e., needed data was not included in the plot or the plot was not presented logarithmically when it should have been). However, this problem did not seem to have prevented the scientists from trying to make the changes: there were no instances of a scientist saying the visualization tool was too complicated or difficult to work with (though these tools could no doubt be improved). Thus, the scientists' use of spatial transformations do not seem to be a substitute for "bad" graphs, but rather a strategy to understand the data more thoroughly.

As suggested earlier, current theories of graph comprehension can not account for this pattern of results. Current theories (e.g., Bertin, 1983; Kosslyn, 1989; Pinker, 1990) deal primarily with how users extract information that is explicitly available on a graph or vi-

sualization. In this study, we have shown that users do not simply extract information that is explicitly shown on a visualization; rather, they extract information and use mental imagery to create similar visualizations, modify those mental images, and compare their mental results to on-screen results. These spatial transformation seem to be used for a variety of reasons, including hypothesis testing and understanding their own mental representation through a process of aligning various mental images (Trafton et al., 2001).

How can we integrate these new results into current theories? We believe that the current theoretical model should be expanded to include spatial transformations as part of the cognitive processes that users go through to interpret and use visualizations. Figure 4 shows our current model of graph comprehension, along with evidence that supports each stage of this model.

We believe, as Figure 4 shows, that when people use graphs or visualizations, they initially go through a process to understand the graph itself. Then, when they need to extract information, they can either read off that information directly from the visualization or, if that information is not available, perform a spatial transformation to get the needed information.[2] Finally, that information is actually used by the user.

Acknowledgments: This research was supported in part by grant 55-7850-00 to the first author from the Office of Naval Research. We would also like to thank Wai-Tat Fu, Anthony Harrison, and William Liles for comments on a previous draft of this paper.

References

Bertin, J. (1983). *Semiology of graphs*. Madison, WI: University of Wisconsin Press.

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Joural of Experimental Psychology: Applied*, *4*(2), 75–100.

Carter, L. F. (1947). An experiment on the design of tables and graphs used for presenting numerical data. *Journal of Applied Psychology*, *31*, 640–650.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: theory, experimentation, and application to the development of graphical method. *Journal of the American Statistical Association*, *79*, 531–553.

Cleveland, W. S., & McGill, R. (1986). An experiment in graphical perception. *International Journal of Man-Machine Studies*, *25*, 491–500.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg, & J. E. Davidson (Eds.), *The nature of insight*, (pp. 365–395). Cambridge, MA: MIT Press.

Dunbar, K. (1996). How scientists think: Online creativity and conceptual change in science. In T. B. Ward, S. M. Smith, & S. Vaid (Eds.), *Creative thought: An Investigation of Conceptual Structures and Processes*, (pp. 461–493). Washington, DC: APA Press.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Revised edition). Cambridge, MA: MIT Press.

Hegarty, M. (1992). Mental animation: Inferring motion from static displays of mechanical systems. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *18*(5), 1084–1102.

Kosslyn, S. M. (1989). Understanding charts and graphs. *Applied Cognitive Psychology*, *3*, 185–226.

Kosslyn, S. M., Sukel, K. E., & Bly, B. M. (1999). Squinting with the mind's eye: Effects of stimulus resolution on imaginal and perceptual comparisons. *Memory and Cognition*, *27*(2), 276–287.

Lohse, G. L. (1993). A cognitive model for understanding graphical perception. *Human Computer Interaction*, *8*, 353–388.

Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), *Artificial intelligence and the future of testing*, (pp. 73–126). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*, 701–703.

Trafton, J. G., Kirschenbaum, S. S., Tsui, T. L., Miyamoto, R. T., Ballas, J. A., & Raymond, P. D. (2000). Turning pictures into numbers: Extracting and generating information from complex visualizations. *International Journal of Human Computer Studies*, *53*(5), 827–850.

Trafton, J. G., Trickett, S. B., & Mintz, F. E. (2001). Overlaying images: Spatial transformations of complex visualizations. Paper to be presented at Model-Based Rasoning: Scientific Discovery, Technological Innovation, Values in Pavia, Italy.

Trickett, S. B., Fu, W., Schunn, C. D., & Trafton, J. G. (2000a). From dipsy-doodles to streaming motions: Changes in representation in the analysis of visual scientific data. In *Proceedings of the Twenty Second Annual Conference of the Cognitive Science Society*.

Trickett, S. B., Trafton, J. G., & Schunn, C. D. (2000b). Blobs, dipsy-doodles and other funky things: Framework anomalies in exploratory data analysis. In *Proceedings of the Twenty Second Annual Conference of the Cognitive Science Society*.

[2]This model does not show the other ways to gather information because it did not show up as a major component in our current datasets.

That s Odd! How Scientists Respond to Anomalous Data

Susan B. Trickett
(stricket@gmu.edu)
Dept. of Psychology
George Mason University
Fairfax, VA 22030 USA

J. Gregory Trafton
(trafton@itd.nrl.navy.mil)
Naval Research Laboratory
NRL Code 5513
Washington, DC 20375

Christian D. Schunn
(schunn@gmu.edu)
Dept of Psychology
George Mason University
Fairfax, VA 22030 USA

Anthony Harrison
(aharris8@gmu.edu)
Dept. of Psychology
George Mason University
Fairfax, VA 22030 USA

Abstract

We use an *in vivo* methodology to investigate the responses of scientists to anomalies. Protocols of 3 scientists performing data analysis in 2 domains were analyzed. We found that the scientists noticed anomalies and paid more attention to them than to expected data. This attention took the form of proposing a hypothesis and then elaborating that hypothesis by reference to other data in the visual display, rather than to the scientists theoretical domain knowledge.

Introduction

How do scientists deal with anomalous or unexpected data? Philosophers of science (e.g., Kuhn, 1962; Lakatos, 1976) have argued that anomalies play a crucial role in moving science forward. Scientists themselves have claimed that investigating anomalies lies at the heart of scientific innovation (e.g., Knorr, 1980).

Psychologists have been interested in the cognition underlying how scientists deal with anomalies. There have been two general approaches to the study of anomalies. One approach focuses on response to negative evidence in concept identification tasks (Wason, 1960). Several studies have found that people are likely to seek confirming evidence for their theories (e.g., Mynatt, Doherty, & Tweney, 1977).

Surprisingly, studies from this tradition have found that scientists are also very susceptible to confirmation bias (e.g., Mahoney & DeMonbreun, 1977). One criticism of this approach is that the tasks are abstractions of the hypothesis-testing cycle and therefore do not allow the participants to make use of their extensive domain knowledge (e.g., Chinn & Malhotra, 2001). However, sociological studies (e.g., interviews) of practicing scientists have also found that scientists appear to display confirmation bias (Mitroff, 1974).

A second approach investigates scientists' response to anomalies as they perform analyses of authentic scientific data. This approach includes both historical studies of scientific discovery (Chinn & Brewer, 1992; Kulkarni & Simon, 1988; Nersessian, 1999) and *in vivo* observations of contemporary practicing scientists (Dunbar, 1997; Trickett, Trafton, & Schunn, 2000). Chinn and Brewer have developed a taxonomy of responses, from ignoring the anomaly to changing the theory. They have found evidence for the whole range of responses in historical records of science. Yet the results of other studies suggest that scientists do pay attention to anomalies. For example, Dunbar (1997) found that individual scientists were quick to discard a hypothesis when faced with results that were inconsistent with it, and Kulkarni and Simon (1988) identified an "attend to surprising result" heuristic as crucial to Hans Krebs' discovery of the urea cycle.

Recently, Alberdi, Sleeman and Korpi (2000) have brought together these two approaches to the study of anomalies in scientific thinking. They conducted a psychological study of expert botanists performing a categorization task in the domain of plant taxonomy. As participants formed hypotheses about the category into which a current set of plants would fall, they were presented with a "rogue," or anomalous, item that belied their expectations. Alberdi and his colleagues found that participants did indeed pay attention to the anomalies. Furthermore, they identified a number of strategies by which participants attempted to resolve the anomalous data. Key among these was the "instantiate" strategy, in which participants searched their theoretical domain knowledge for a new hypothesis that would accommodate the anomaly.

Although categorization is an important task in many areas of science, there are many other situations in which scientists might encounter anomalies. For example, to name a few, experimental results might fail to match a prediction, a computational model might yield a different output from expectation, or empirical data might contain puzzling phenomena hard to explain by means of current theoretical understanding. Thus, many questions remain about other circumstances under which a scientist encounters an anomaly in the course of his or her own research.

The goal of this paper is to investigate scientists' response to anomalies they discover as they analyze their own data. Our first question concerns the extent to which scientists attend to anomalous data. One can imagine a range of possible responses, from ignoring the anomaly to attempting to give a full accounting of it (Chinn and Brewer, 1992). Do practicing scientists tend to ignore anomalies as suggested by Mitroff (1974) or do they attend to anomalies as suggested by Dunbar (1997) and Alberdi et al. (2000)? Our second question concerns the processes and strategies by which scien-

tists deal with anomalies when they encounter them. Do they take a theoretical approach, as found by Alberdi et al (2000) or focus on the data itself? Alternatively, additional strategies might emerge from observing scientists at work. We investigate all these possibilities.

Method

In order to investigate the issues discussed above, we have adapted Dunbar s in vivo methodology (Dunbar, 1997; Trickett, Trafton & Schunn, 2000). This approach offers several advantages. It allows observation of experts, who can use their domain knowledge to guide their strategy selection. It also allows the collection of "on-line" measures of thinking, so that the scientists' thought processes can be examined as they occur. Finally, the tasks the scientists do are fully authentic.

Two sets of scientists were videotaped while conducting their own research. All the scientists were experts, having earned their Ph.D.s more than 6 years previously. In the first set, two astronomers, one a tenured professor at a university, the other a fellow at a research institute, worked collaboratively to investigate computer-generated visual representations of a new set of observational data. At the time of this study, one astronomer had approximately 20 publications in this general area, and the other approximately 10. The astronomers have been collaborating for some years, although they do not frequently work at the same computer screen and the same time to examine data.

In the second dataset, a physicist with expertise in computational fluid dynamics worked alone to inspect the results of a computational model he had built and run. He works as a research scientist at a major U.S. scientific research facility and had earned his Ph.D. 23 years ago. He had inspected the data earlier but made some adjustments to the physics parameters underlying the model and was therefore revisiting the data.

Both sets of scientists were instructed to carry out their work as though no camera were present and without explanation to the experimenter (Ericsson & Simon, 1993). The relevant part of the astronomy session lasted about 53 minutes, and the physics session, 15 minutes. All utterances were later transcribed and segmented according to complete thought. All segments were coded by 2 coders as on-task (pertaining to data analysis) or off-task (e.g., jokes, phone interruptions, etc.). Inter-rater reliability for this coding was more than 95%. Off-task segments were excluded from further analysis. On-task segments (N = 649 for astronomy and N = 176 for physics) were then grouped into episodes (N = 19 for astronomy and N = 9 for physics). Episodes began with the scientists' focus on a phenomenon and lasted until attention switched to another feature or theoretical issue. This grouping of the protocol into episodes allowed us to focus on the more immediate reaction to anomalies.

The Tasks and the Data

Astronomy The data under analysis were optical and radio data of a ring galaxy. The astronomers high-level goal was to understand its evolution and structure by understanding the flow of gas in the galaxy. In order to understand the gas flow, the astronomers must make inferences about the velocity field, represented by contour lines on the 2-dimensional display.

The astronomers task was made difficult by two characteristics of their data. First, the data were one- or at best 2-dimensional, whereas the structure they were attempting to understand was 3-dimensional. Second, the data were noisy, with no easy way to separate noise from real phenomena. Figure 1 shows a screen snapshot of the type of data the astronomers were examining. In order to make their inferences, the astronomers used different types of image, representing different phenomena (e.g., different forms of gas), which contain different information about the structure and dynamics of the galaxy. In addition, they could choose from images created by different processing algorithms, each with advantages and disadvantages (e.g., more or less resolution). Finally, they could adjust some features of the display, such as contrast or false color.

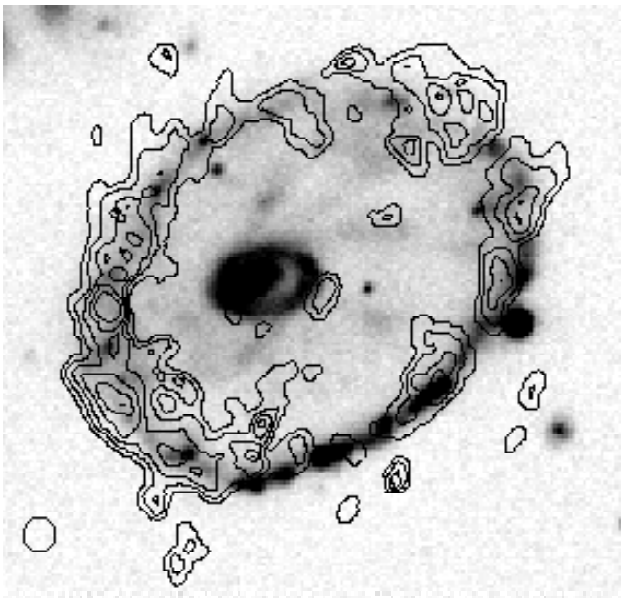


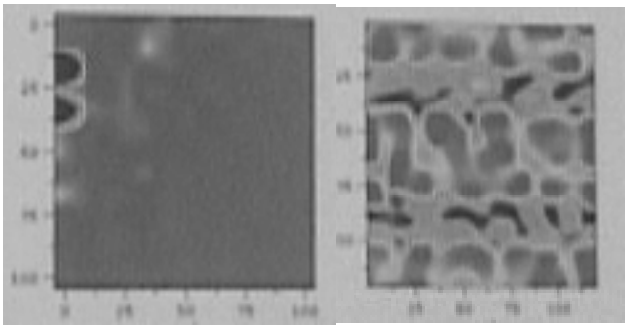
Figure 1: Example of data examined by astronomers. Radio data (contour lines) are laid over optical data.

Physics The physicist was working to evaluate how deep into a pellet a laser light will go before being reflected. His high-level goal was to understand the fundamental physics underlying the reaction, an understanding that hinged on comprehending the relative importance and growth rates of different modes. The physicist had built a model of the reaction; other scientists had independently conducted experiments in which lasers were fired at pellets and the reactions recorded. A

close match between model and empirical data would indicate a good understanding of the underlying theory. Although the physicist had been in conversation with the experimentalist, he had not viewed the empirical data, and in this session he was investigating only the results of his computational model. However, he believed the model to be correct (i.e., he had strong expectations about what he would see), and in this sense, this session may be considered confirmatory.

The data consisted of two different kinds of representation of the different modes, shown over time (nanoseconds). The physicist was able to view either a Fourier decomposition of the modes or a representation of the raw data. Figure 2 shows an example of the physicist's data. He could choose from black-and-white or a variety of color representations, and could adjust the scales of the displayed image, as well as some other features. He was able to open numerous views simultaneously. A large part of his task was comparing images, both different types of representation of the same data and different time slices represented in the same way.

Figure 2: Example of data examined by physicist
Fourier modes (left) and raw data (right)



Coding Scheme

Our goal in this research was to investigate scientists' response to anomalous data. First, we wanted to establish whether and to what extent the scientists noticed and attended to anomalies. Second, we wanted to investigate the processes by which they respond.

Both protocols were coded independently by 2 different coders. Initial inter-rater reliability for each code was greater than 85%. Disagreements were resolved by discussion. Any coding disagreements that could not be resolved were excluded from further analysis.

Noticings In order to establish which phenomena unusual or not the scientists attended to, we first coded for the scientists' *noticing* phenomena in the data. A noticing could involve merely some surface feature of the display, such as a line, shape, or color, or it could involve some interpretation, for example, identifying an area of star formation or the implied presence of a mode. Only the first reference to a phenomenon was coded as a noticing; coding of subsequent references to the same phenomenon is discussed below.

Because our investigation focused on the extent to which the scientists attended to anomalies in the data, we further coded these noticings as either "anomalous" or "expected," according to one or more of the following criteria: a) in some cases the scientist made explicit verbal reference to the fact that something was anomalous or expected; b) if there was no explicit reference, domain knowledge was used to determine whether a noticing was anomalous or not;[1] c) a phenomenon might be associated with (i.e., identified as like) another phenomenon that had already been established as anomalous or not; d) a phenomenon might be contrasted with (i.e., identified as unlike) a phenomenon that had already been established as anomalous or not; e) a scientist might question a feature, thus implying that it is unexpected. Table 1 illustrates these codes.

Criterion	Code	Example
Explicit	Anomalous	What's *that funky thing* That's odd
Domain Knowledge	Expected	You can see that *all the H1* is concentrated in the ring
Association	Anomalous	You see *similar kinds of intrusions* along here
Contrast	Expected	That's odd As opposed to *these things*, which are just the lower contours down here
Question	Anomalous	I still wonder why *we don't see any H1 up here* in this sort of northern ring segment?

Table 1: Noticings (in italics): anomalous or expected

Subsequent References One of our questions was the extent to which the scientists attended to anomalies. The coding of noticings captured only the first reference to a phenomenon of interest; we needed to establish how frequently they made subsequent reference to each noticing. Consequently, all subsequent references were also identified and coded.[2] Not all subsequent references immediately followed a noticing; frequently, the scientists returned to a phenomenon after investigating other features of the data. Subsequent references were identified both within the episode in which the noticing had occurred and across later episodes.

The rest of the coding scheme addresses *how* the scientists responded to the anomalies, in particular imme-

[1] The coders domain knowledge came from textbooks and interviews with the scientists.

[2] In the astronomy dataset, because the scientists shared a computer monitor, frequently the first interaction between them after a noticing was to make sure they were both looking at the same thing. Subsequent references that served purely to establish identity were *not* included in the analyses.

diately after they notice the anomalies. To investigate the scientists immediate response to their anomalous findings, we coded 10 utterances following each noticing, whether anomalous or expected (minus utterances establishing which phenomenon was under discussion, in the astronomy dataset). We anticipated that scientists would attempt to produce hypotheses for the anomalies, and that some of these hypotheses would be discussed further. Based on the results reported by Alberdi, et al. (2000), we investigated the extent to which elaboration of hypotheses was grounded in theory or in the visual display of the data. We also anticipated the use of additional strategies and inspected the data to identify strategies that emerged, as discussed below.

Hypotheses Statements that attempted to provide a possible explanation for the data were coded as hypotheses. All hypotheses were further coded as *elaborated* or *unelaborated*. Elaboration consisted of one or more statements that either supported or opposed the hypothesis. Hypotheses that were not discussed after they were proposed were coded as unelaborated.

When a hypothesis was elaborated, we coded whether the elaboration was *theoretical* or *visual*. When evidence for or against a hypothesis was grounded in theoretical domain knowledge, elaboration was coded as theoretical; when evidence came from the display, it was coded as visual.

Place in context A strategy that emerged from our examination of the data was considering the noticing in relation to other data. Thus we coded whether or not the scientist placed the noticing in context, and whether that context was another part of the dataset (*local*) or the scientist's own theoretical knowledge (*global*).

Results and Discussion

Noticing Anomalies

Our first question was did the scientists notice anomalies in the data?[3] Recall that a noticing is a first-time reference to a phenomenon of interest. Table 2 presents the total number of noticings for each dataset and the percentages of anomalous and expected phenomena. As Table 2 shows, at least one-third of the phenomena the astronomers identified and almost one-half the physicist identified were unusual in some way. It appears then that the scientists *did* notice anomalies in their data.

	Total Noticing	Anomalous	Expected	Not coded
Astronomy	27	33%	48%	19%
Physics	9	44%	44%	12%

Table 2: Frequency of anomalies and expected noticings

[3] We presented a more detailed discussion of a subset of the results for the astronomy dataset in Trickett et al. (2000).

Attention to Anomalies

Once the scientists had identified something unusual in the data, what did they do with this observation? There are several possible reactions: they could pursue the anomaly in order to try to account for it, they might temporarily disregard it but return to it later, or they might move on to explore some other, better understood, aspect of the data. A related question is whether their response to anomalies was different from their response to expected phenomena.

We investigated this issue by counting how often the scientists made subsequent reference to a noticing immediately upon identifying it. If anomalies and expected phenomena are of equal interest, we would expect them to make a similar number of references to both the anomalous and expected patterns. However, if anomalies play a more important role in their efforts to understand the data, we would expect them to pay more attention (measured by the number of subsequent references) to anomalies than to expected observations.

As Table 3 shows, for both the astronomy and physics datasets, scientists paid more attention to anomalies than expected phenomena, $t(28)=2.33$, $p<.05$. In the case of astronomy, the anomalies received over 3 times as many subsequent references within the same episode as the expected phenomena. The physics dataset follows a similar pattern, with more than twice as many references to anomalies as expected phenomena. The results are in stark contrast to the findings of the confirmation bias literature.

	Anomalies	Expected
Astronomy	7.6	1.5
Physics	3.0	1.25

Table 3: Mean number of subsequent references per noticed object to anomalies and expected phenomena

Immediate Response to Anomalies

We have shown that when the scientists noticed an anomaly, they immediately attended to it, but we have not analyzed the content of that attention to anomalies. In order to understand what kind of the scientists made, we now turn to the results of the second part of our coding scheme, which was applied to the 10 utterances that immediately followed the initial noticing of anomalies and expected phenomena.

Identify Features As Figure 3 shows, the scientists were only slightly (and nonsignificantly) more likely to identify specific features of the anomalies as the expected noticings, and this pattern held for both domains.

Propose Hypothesis As Figure 4 shows, the scientists were much more likely to propose a hypothesis for the

anomalies than the expected noticings $\chi^2(1) = 7.5$, $p<.05$, and this pattern was very strong in both domains.

Figure 3: Percentage of noticings for which scientists identified features.

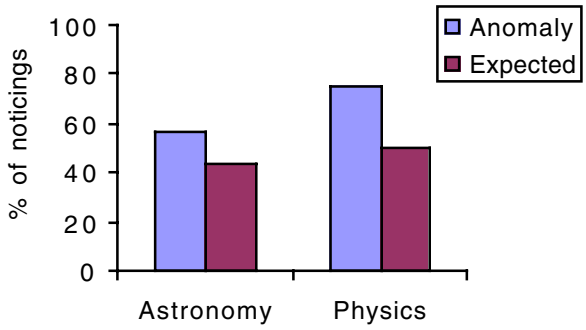
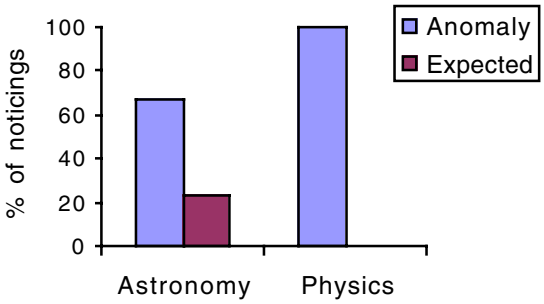
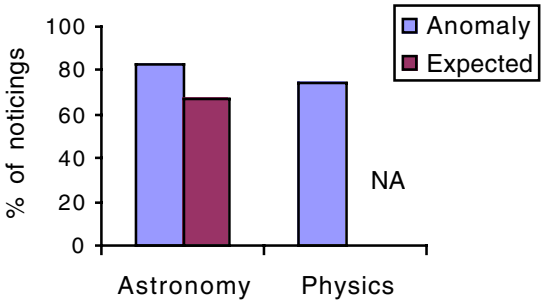


Figure 4: Percentage of noticings with hypotheses



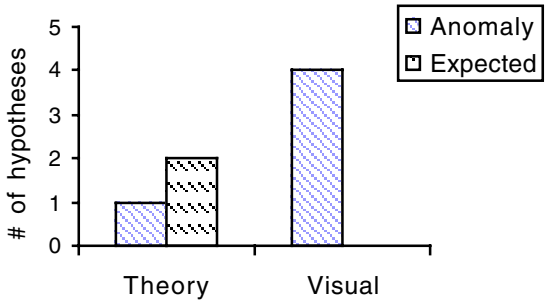
Elaborated Hypothesis Once the scientists had proposed a hypothesis (primarily about the anomalies), in most cases they elaborated that hypothesis. Figure 5 presents the proportion of hypotheses that were elaborated within each domain for expected and anomalous noticings. In most cases, scientists attempted to elaborate the hypotheses, for both expected and anomalous noticings (note that there were no hypotheses to elaborate in the expected physics case).

Figure 5: Percentage of noticings that were elaborated



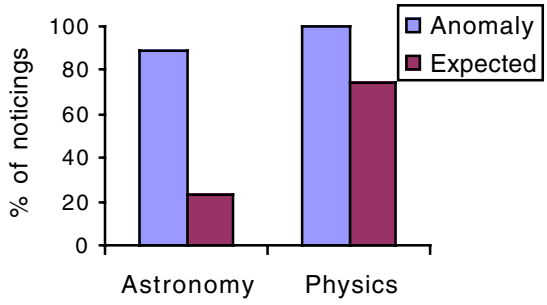
Source of Elaboration For the physics dataset, there were not enough elaborated hypotheses to analyze further. For the astronomy data, evidence about 4 of the 5 hypotheses about anomalies came from the visual display. The 2 hypotheses about expected noticings were developed theoretically. Figure 4 shows this result.

Figure 6: Elaboration type for hypotheses (astronomy)



Place in Context In addition to (or instead of) developing hypotheses about the noticings, the scientists also might consider the noticing in relation to other information, either theoretical information in memory (global context) or information about the current dataset (local context), or they might not place it in either context. In fact, none of the noticings was considered in the context of the scientists' theoretical knowledge (global). However, the scientists considered the noticings in the context of the current dataset (local), and this sequence occurred more frequently for the anomalies than for the expected phenomena, especially in the astronomy dataset (see Figure 7), $\chi^2(1) = 9.21$, $p<.01$.

Figure 7: Percentage of noticings put in local context



General Discussion and Conclusion

We examined the behavior of scientists at work, analyzing their own data. Our results show that these scientists not only notice anomalies in the data, but also attend to them, contrary to the confirmation bias literature, but similar to the findings of Dunbar (1997) and Alberdi et al. (2000).

The scientists we observed not only notice and attend to anomalies, but also do so in a particular way. Furthermore, this pattern is quite different from the pattern that results from their observation of expected phenomena. When they notice an expected phenomenon, after identifying or describing its features, the scientists are likely to engage in no further elaboration of the phenomenon. On the rare occasions when they do attempt to account for it by proposing a hypothesis, they seek evidence in their own theoretical knowledge, rather than in the visually displayed data. By contrast, however, for anomalous noticings, the scientists attempt to account for the anomaly by proposing a hypothesis. They then elaborate the hypothesis, primarily by seeking evidence in the visual display, and finally consider how the anomaly relates to neighboring phenomena within the same display.

Our results mesh in part with those of other researcher, in that they provide further evidence for the important role played by anomalies as scientists analyze and reason about data. However, our results differ from those of Alberdi et al. (2000) in some significant ways. When the botanists in their study encountered an anomaly, they were most likely to use a strategy of theory-driven search for an explanation. The scientists in our study, however, sought support for hypotheses in the visually displayed data, and attempted to place the anomaly in the local context of neighboring phenomena. Only hypotheses about expected phenomena were developed at a theoretical level.

There are several possible explanations for this difference. Situational differences in the tasks performed by the participants in these two studies might affect their strategy. For the botanists, categorization was the goal *per se.* Although the astronomers and physicist were performing some categorization tasks, this was done in service of understanding the data as a whole, in order to build a mechanistic theory. The difference in their goals might account for the different strategies they used. Another possibility is that the botanists were getting immediate feedback on their hypotheses, whereas the other scientists had to generate their own feedback. In this sense, the botanists' task is similar to a supervised learning task, whereas the astronomers and physicist were in a situation where learning was unsupervised (Hertz, Krogh, & Palmer, 1991). It is plausible that the uncertainties inherent in this situation can account for the fact that these scientists sought feedback in the empirical data in the display rather than jumping immediately to their theoretical domain knowledge.

Acknowledgments

This research was supported in part by grant 55-7850-00 to the second author from ONR. We thank Wai-Tat Fu and William Liles for comments.

References

Alberdi, E., Sleeman, D. H. & Korpi, M. (2000). Accommodating surprise in taxonomic tasks: The role of expertise. *Cognitive Science, 24*(1), 93-122.

Chinn, C. A., & Brewer, W. F. (1992). Psychological reponses to anomalous data. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Chinn, C. A., & Malhotra, (2001). Epistemologically authentic scientific reasoning. In K. Crowley, C. D. Schunn & T. Okada, (Eds.), *Designing for science: Implications from everyday, classroom, and professional settings.* Mahwah, NJ: Erlbaum.

Dunbar, K. (1997). How scientists think: On-line creativity and conceptual change in science. In T. B. Ward, S. M. Smith, & J. Vaid, (Eds.), *Creative thought: An investigation of conceptual structures and processes*. Washington, DC, USA: APA Press.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Hertz, J., Krogh, A., and Palmer, R. G. (1991). Introduction to the theory of neural computation. Addison-Wesley, Reading, MA.

Kuhn, T. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Lakatos, I. (1976). *Proofs and refutations.* Cambridge, UK: Cambridge University Press.

Kulkarni, D., & Simon, H. A. (1988). The processes of scientific discovery: The strategy of experimentation. *Cognitive Science, 12*(2), 139-175.

Mahoney, M. J., & DeMonbreun, B. G. (1977). Psychology of the scientist: An analysis of problem-solving bias. *Cognitive Therapy and Research, 3,* 229-238.

Mitroff, I. (1974). *The subjective side of science: A philosophical inquiry into the psychology of the Apollo moon scientists.* Amsterdam: Elsevier.

Mynatt, C. R. , Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology, 29*(1), 85-95.

Nersessian, N. (1999). Model based reasoning in conceptual change. In L. Magnani & N. Nersessian (Eds.), *Model-based reasoning in scientific discovery* (pp. 5-22). New York, NY: Kluwer Academic.

Trickett, S. B., Trafton, J. G., & Schunn, C. D. (2000). Blobs, dipsy-doodles and other funky things: Framework anomalies in exploratory data analysis. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12,* 129-140.

Spoken Language Comprehension Improves the Efficiency of Visual Search

Melinda J. Tyler (mjt15@cornell.edu)
Department of Psychology, Cornell University
Ithaca, NY 14853 USA

Michael J. Spivey (spivey@cornell.edu)
Department of Psychology, Cornell University
Ithaca, NY 14853 USA

Abstract

Much recent eye-tracking research has demonstrated that visual perception plays an integral part in on-line spoken language comprehension, in environments that closely mimic our normal interaction with our physical environment and other humans. To test for the inverse, an influence of language on visual processing, we modified the standard visual search task by introducing spoken linguistic input. In classic visual search tasks, targets defined by only one feature appear to "pop-out" regardless of the number of distractors, suggesting a parallel search process. In contrast, when the target is defined by a conjunction of features, the number of distractors in the display causes a highly linear increase in search time, suggesting a more serial search process. However, we found that when a conjunction target was identified by a spoken instruction presented concurrently with the visual display, the effect of set size on search time was dramatically reduced. These results suggest that the incremental linguistic processing of the two spoken target features allows the visual search process to, essentially, conduct two nested single-feature parallel searches instead of one serial conjunction search.

Introduction

For a psycholinguist studying spoken language comprehension, the visual environment would be considered "context". However, for a vision researcher, the visual environment is the primary target of study, and auditory/linguistic information would be considered the "context". Clearly, this variable use of the label "context" is due to differences in perspective, not due to any objective differences between language and vision. In everyday perceptual/communicative circumstances, humans must integrate visual and linguistic information extremely rapidly for even the simplest of exercises. Consider the real-time dance of linguistic, visual, and even gestural events that takes place during a conversation about the weather. This continuous coreferencing between visual and linguistic signals may render the very idea of labeling something as "context" arbitrary at best, and perhaps even misleading.

The problem of "context" has traditionally been dealt with in a rather drastic fashion: researchers forcibly ignore it. If context does not influence the primary functions of the process of interest (be it in language, vision, memory, reasoning, or action), then that process can be thought of as an encapsulated module which will permit dissection via a nicely limited set of theoretical and methodological tools. For example, prominent theories of visual perception and attention posit that the visual system is functionally independent of other cognitive processes (Pylyshyn, 1999; Zeki, 1993). This kind of modularity thesis has been applied to accounts of language processing as well (Chomsky, 1965; Fodor, 1983). As a result, a great deal of progress has been made toward developing first approximations of how vision may function and how language may function.

However, recent eye-tracking studies have shown evidence that visual perception constrains real-time spoken language comprehension. For example, temporary ambiguities in word recognition and in syntactic parsing are quickly resolved by information in the visual context (Allopenna, Magnuson, & Tanenhaus, 1998; Spivey & Marian, 1998; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Findings like these are difficult for modular theories of language to accommodate.

The present experiment demonstrates the converse: that language processing can constrain visual perception. In a standard visual search task, a target object is typically defined by a conjunction of features, and reaction time increases linearly with the number of distractors, often in the range of 15-25 milliseconds per item (Duncan & Humphreys, 1989; Treisman & Gelade, 1980; Wolfe, 1994). However, when we presented the visual display first, and then provided the spoken target features incrementally, we found that reaction time was considerably less sensitive to the number of distractors.

With conjunction search displays, increased reaction times as a linear function of set size were originally interpreted as evidence for serial processing of the objects in the display, and contrasted with the near-flat function of reaction time by set size observed with

feature search displays -- where a single feature is sufficient to identify the target object. It was argued that the early stages of the visual system process individual features independently and in parallel (Livingstone & Hubel, 1988), allowing the target object to "pop out" in the display if it is discriminable by a single feature, but requiring application of an attentional window to the individual objects, one at a time, if the target object is discriminable only by a conjunction of features (Treisman & Gelade, 1980). This categorical distinction between parallel search of single feature displays and serial search of conjunction displays has been supported by PET scan evidence for a region in the superior parietal cortex that is active during conjunction search for motion and color, but not during single feature search for motion or for color (Corbetta, Shulman, Miezin, & Petersen, 1995).

However, several studies have discovered particular conjunctions of features that do not produce steeply sloped reaction-time functions by set size (e.g., McLeod, Driver & Crisp, 1988; Nakayama & Silverman, 1986). Additionally, it is possible to observe the phenomenology of 'pop-out' while still obtaining a significant (albeit, small) effect of set size on reaction time (Bridgeman & Aiken, 1994). Moreover, it has been argued that steeply sloped reaction-time functions may not reflect serial processing of objects in the display, but rather noise in the human visual system (Eckstein, 1998; Palmer, Verghese, & Pavel, 2000). Overall, a wide range of studies have suggested that the distinction between putatively "serial" and "parallel" search functions is continuous rather than discrete, and should be considered extremes on a continuum of search difficulty (Duncan & Humphreys, 1989; Nakayama & Joseph, 1998; Olds, Cowan, Jolicoeur, 2000; Wolfe, 1994, 1998).

In a recent study, Spivey, Tyler, Eberhard, and Tanenhaus (in press b) demonstrated that the incremental processing of linguistic information could, essentially, convert a difficult conjunction search into a pair of easier searches. When target identity was provided via recorded speech presented concurrently with the visual display, displays that typically produced search slopes of 19 ms per item produced search slopes of 8 ms per item. It was argued that if a spoken noun phrase such as "the red vertical" is processed incrementally (cf. Altmann, & Kamide, 1999; Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995; Marslen-Wilson, 1973, 1975), and there is extremely rapid integration between partial linguistic and visual representations, then one might predict that the listener should be able to search items with the first-mentioned feature before even hearing the second one. If the observer can immediately attend to the subset of objects sharing that first-mentioned feature, such as the target color (Egeth, Virzi, & Garbart, 1984; Friedman-Hill & Wolfe, 1995; Motter & Holsapple, 2000), and subsequently search for the target object in that subset upon hearing the second-mentioned feature, then this initial immediate group selection should reduce the effective set size to only those objects in the display that share the first-mentioned feature – effectively cutting the search slope in half.

At least two concerns remain before this basic finding can be extended and tested in the many different variations of visual search displays. First, since a slope of 8 ms per item is clearly in the range of what has traditionally been considered “parallel search”, it is somewhat unclear whether the result is in fact a *halving* of the effective set size or a near *elimination* of the effect of set size. Essentially, the question is whether the first feature extraction is a genuine “pop-out” effect and the second is a genuine serial search of those “popped out” objects (half of the set size), or are both searches “practically parallel”. A replication of the study may provide some insight into this question. Second, the experiments reported by Spivey et al. (in press b) ran participants in separate blocks of control trials and trials with concurrent auditory/visual input. It is in principle possible that practice was somehow more effective in the auditory/visual concurrent condition, or that subjects developed some unusual strategy in that condition that they didn’t use in the control condition. To be confident in the result, it is necessary to replicate it with a mixed (instead of blocked) design, where the control trials and the A/V concurrent trials are randomly interspersed.

Experiment

Method

Participants Eighteen Cornell undergraduate students were recruited from various Psychology classes. Participants were reimbursed 1 point of course extra credit for participating in the study.

Procedure The experiment was composed of two types of trials presented in random mixed order within one continuous block of 192 trials. Participants were instructed to take breaks between trials when they felt it was necessary. In one type of trial, the participant was auditorily informed of the target identity *before* presentation of the visual display (‘Auditory First’ control condition). In the other type of trial, the participant was auditorily informed of the two defining feature words of the target *concurrently with* the onset of the visual display (‘A/V Concurrent’ condition) (see Figure 1) Of the 192 trials, 96 were ‘Auditory First’, and 96 were ‘A/V concurrent.’

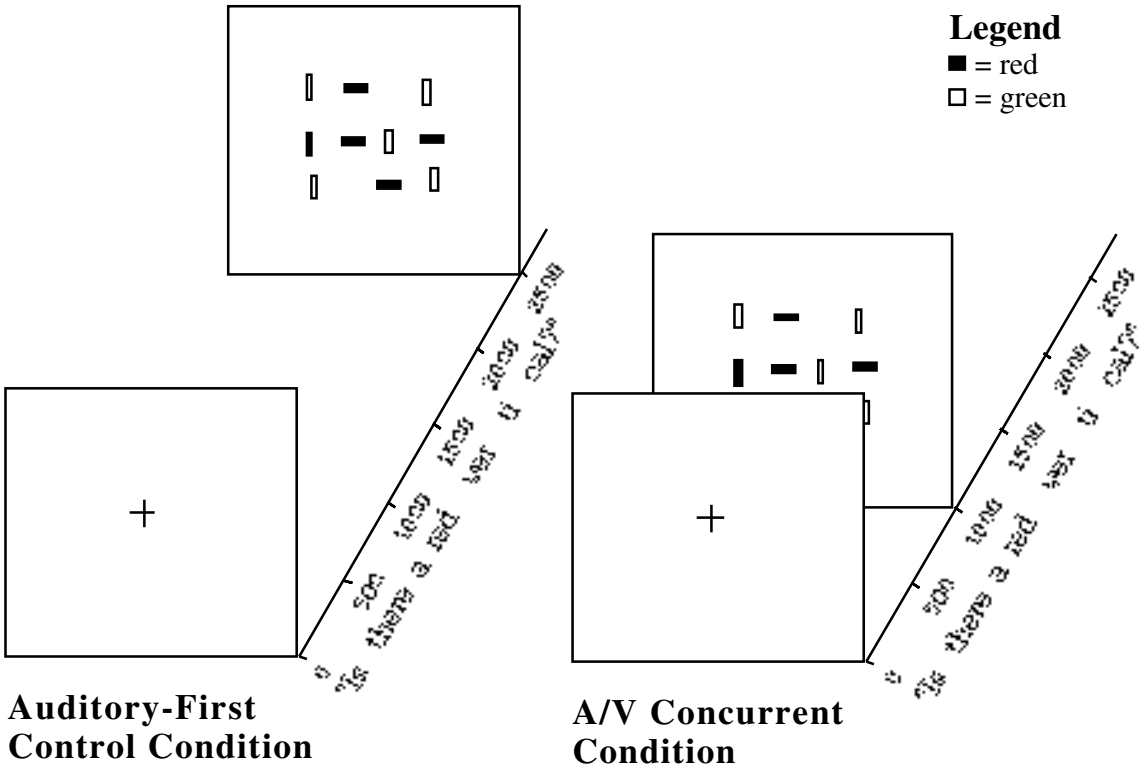


Figure 1. Schematic diagram of the two conditions. In the Auditory-First condition, the search display is presented after the entire spoken query is heard, whereas in the A/V Concurrent condition, the search display is presented immediately before the two target features are heard. Reaction time is measured from the point of display onset.

Trials began with a question delivered in the format of a speech file. The same female speaker recorded all speech files with the same preamble recording, "Is there a…" being spliced onto the beginning of each of the four types of target query types ("…red vertical?", "…red horizontal?", "…green vertical?", and "…green horizontal?"). Each of the four types of speech files were edited to be almost identical in length, and with almost identical auditory spacing of defining feature words. Participants were instructed to press a 'yes' key on a computer keyboard if the queried object was present in the display, and the 'no' key if it was absent. It was stressed to participants that they should do this as quickly and accurately as possible. An initial fixation cross preceded the onset of the visual display in order to direct participants' gaze to the central region of the display. Each stimulus bar subtended 2.8 degrees X 0.4 degrees of visual angle, and neighboring bars were separated from one another by an average of 2 degrees of visual angle. Trials with red vertical bars as targets and trials with green vertical bars as targets, as well as red and green horizontal bars as targets, were equally and randomly distributed throughout the session. All participants had normal or corrected-to-normal vision, and all had normal color perception. The objects comprising the visual display appeared in a grid-like arrangement positioned centrally in the screen (see Figure 1). Set sizes of objects comprising the visual displays were 5, 10, 15, and 20.

Results

Mean accuracy was 95% and did not differ across conditions. Figure 2 shows the reaction time by set size functions for target-present trials (filled symbols) and target-absent trials (open symbols) in the A/V Concurrent condition and the Auditory-First condition. The best-fit linear equations are accompanied by their r^2 values indicating the percentage of variance accounted for by the linear regression.

Overall mean reaction time was slower in the A/V Concurrent condition as a result of the complete auditory notification of target identity being delayed by approximately 1.5 seconds relative to the Auditory-First control condition. However, since spoken word recognition is incremental, participants were able to begin processing before both target feature words had been presented, and overall reaction time was only delayed by about 600 milliseconds.

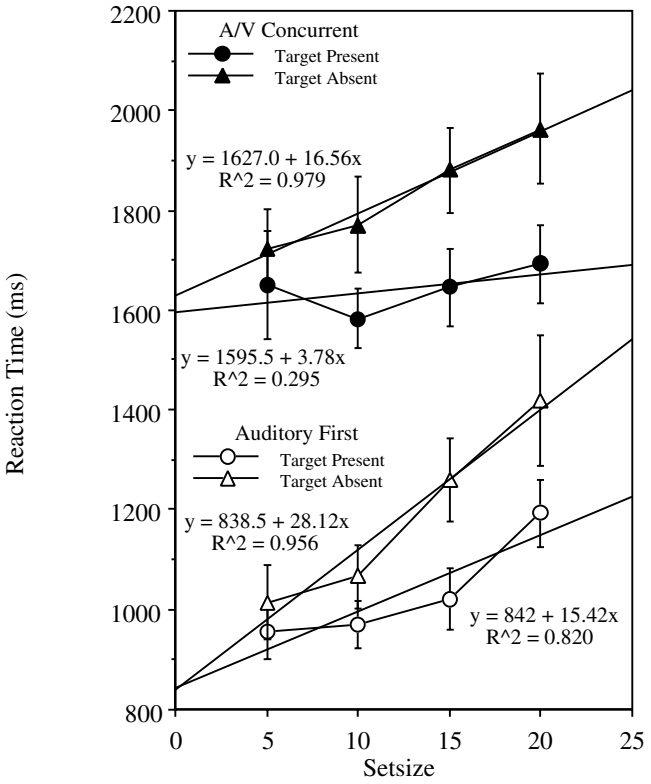


Figure 2: Reaction time as a function of set size.

Repeated-measures analysis of variance revealed significant main effects of Condition [F(1, 16)=230.27, p<.01], Target Presence/Absence [F(1,16)=27.97, p<.01], and Set Size [F(3, 48)=22.83, p<.01]. The effect of Condition was simply that overall reaction times were slower when the delivery of target identity was delayed in the A/V Concurrent condition. The effect of Target Presence/Absence was the common finding that it takes longer to determine that the target is absent than to determine that it is present. Essentially, the target-present case involves something akin to a self-terminating search, and the target-absent case requires something like an exhaustive search. The main effect of Set Size simply showed that, when Condition and Target Presence/Absence are ignored, having more distractors increased reaction time. There was also an interaction between Set Size and Target Presence/Absence, showing that the effect of Set Size was more pronounced in the target-absent trials than in the target-present trials [F(3,48)=4.36, p<.01].

The important result for the purposes of our inquiry was the significant interaction between Condition and Set Size, indicating that the effect of Set Size was more pronounced in the Auditory-First control condition than in the A/V Concurrent condition [F(3, 48)=4.92, p<.01]. Despite the fact that the visual displays were identical, results indicated shallower slopes for reaction-time functions in the A/V Concurrent condition compared to the Auditory-First control condition (Figure 2).

To specifically test whether the mean slope was significantly shallower in the A/V Concurrent condition, participants' individual set size slopes from the two conditions were compared via paired t-tests, and revealed significantly shallower slopes for the A/V Concurrent condition in target-present trials [3.78 vs. 15.42 ms per item; t(16)=2.09, P<.05] and in target-absent trials [16.56 vs. 28.12 ms per item; t(16)=3.39, P<.01]. These results indicate that adjusting the timing of the spoken query (e.g., "Is there a red vertical?"), such that the two target feature words were presented while the visual display was visible, allowed participants to find the target object in a manner that was substantially less affected by the total number of distractors. In fact, the mean slope of 3.78 ms per item in the target-present trials of the A/V Concurrent condition was not significantly greater than zero [t(16)=1.12, p>.25], whereas the mean slope of 15.42 ms per item in the target-present trials of the Auditory-First control condition was robustly greater than zero [t(16)=4.47, p<.001].

Hence it appears that, in the Auditory-First condition, the search process may employ a conjunction template to find the target, thus forcing a serial-like process akin to sequentially comparing each object to the target template. However, in the A/V concurrent condition, it appears that the incremental nature of the speech input allows the search process to begin when only a single feature of the target identity has been heard. This initial single-feature search proceeds in a more parallel fashion (with the second-mentioned target feature being used to find the target amidst the attended subset), thus dramatically improving the efficiency of search.

Discussion

The results suggest that, due to the incremental nature of spoken language comprehension (Allopenna et al., 1998; Altmann & Kamide, 1999; Cooper, 1974; Eberhard et al., 1995; Marslen-Wilson, 1973, 1975; Tanenhaus et al., 1995) the listener/observer can selectively attend to the subset of objects that exhibit the target feature which is mentioned first in the speech stream. Upon hearing even just a portion of the second-mentioned target feature a few hundred milliseconds later, the observer can then locate the conjunction target object amidst this attended (spatially noncontiguous) subset. These results highlight the incremental processing of spoken language comprehension, and demonstrate the human brain's ability to seamlessly cross-index partial linguistic representations (of a noun phrase, for example) with partial visual representations (of a cluttered visual display).

A more detailed question remains, concerning whether the improved efficiency in visual search is due to the first-mentioned target feature initiating an instantaneous parallel search and the second-mentioned target feature initiating a serial search among the attended items (thus cutting the search slope in half) or to both spoken target features initiating parallel searches (causing the search slope to look like that of a single-feature search). In Spivey et al. (in press b, Experiments 1 and 2), the target-present search slopes of 7.7 and 8.9 ms per item in the A/V Concurrent conditions were roughly consistent with both of those interpretations. When parallel and serial search were conceived of as discrete categories, any set-size function of less than 10 ms per item was generally interpreted as "parallel search." However, Spivey et al.'s target-present Auditory First conditions produced search slopes of 19.8 and 18.6 ms per item -- approximately twice the A/V Concurrent slopes.

The present results, with a target-present search slope of 3.78 ms per item in the A/V Concurrent condition, appear to support the "two parallel searches" alternative. However, in all likelihood, the two alternatives outlined above rely too much on the discrete distinction between "parallel" and "serial" search. If there is indeed a continuum of search efficiency (Duncan & Humphreys, 1989; Nakayama & Joseph, 1998; Olds, Cowan, Jolicoeur, 2000; Wolfe, 1994, 1998), and conjunction search is not quite accurately described as an object-by-object serial process (Eckstein, 1998; Palmer, Verghese, & Pavel, 2000), then it might be safest to conclude that each spoken target feature initiates a "relatively efficient" search that is not quite parallel and not quite serial. Importantly, the second search appears to be able to work selectively on the output of the first one -- compelling evidence for the continuous incrementality with which linguistic and visual processing can coordinate.

Until now, there has been little or no evidence for real-time visual perception being influenced by linguistic context. However, there is considerable work reporting demonstrations of real-time language processing being influenced by visual context (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; McGurk & MacDonald, 1976; Spivey & Marian, 1998; Tanenhaus et al., 1995). Recent eyetracking research has shown in a number of circumstances that the resolution of temporary ambiguities is immediately biased by information in the visual array. For example, when participants were instructed to "pick up the candy", they often looked first at a *candle* before then fixating the candy (Tanenhaus et al., 1995). A precise timing analysis of the eye movements suggested that, when the candle and candy are in the visual display, participants had mental representations for both 'candle' and 'candy' equally partially active during the first couple hundred milliseconds of the word (Allopenna et al., 1998). When only the candy was in the display, the eye-movement data suggested that word recognition was faster, involving less competition from partially active alternatives.

Similar findings were reported for syntactic ambiguity resolution. When the visual context contained a referential ambiguity (e.g., two apples for the instruction "Put the apple…") that was best resolved by pursuing the correct parse of a syntactic ambiguity ("Put the apple on the towel in the box."), participants eye-movement patterns suggested a fast and correct interpretation of the instruction. When there was no such referential ambiguity (e.g., only one apple), participants produced eye-movement patterns indicating a mis-parse of the instruction (Spivey, Tanenhaus, Eberhard, & Sedivy, in press a; Tanenhaus et al., 1995).

Those effects of visual context immediately influencing language comprehension were initially met with considerable resistance from a substantial portion of psycholinguists. However, when we would discuss those findings with vision researchers, they were often appreciative but not terribly impressed. To them, it made perfect sense that the visual system was important and powerful enough to occasionally tell the language system what to do very quickly. We are curious to see the reaction of the vision research community to the present results.

Returning to our discussion of the notion of "context", which began this paper, it seems that the rapidity with which the visual system and the language system can coordinate their representations suggests that any attempt to label some signal as "context" is doomed to be an arbitrary choice – a choice that risks marginalizing important information sources as well as opaquely lumping discriminable information sources. Essentially, the less we assume encapsulated modular processes in language and in vision, the less use we have for the notion of "context" in language and in vision. Instead of visual processing and linguistic processing, perhaps a considerable portion of our daily mental life is made up of visuolinguistic processing.

Since the human brain is neither a psycholinguist nor a vision researcher (indeed, it is much more than even the two of them combined), it is not susceptible to the biased perspectives they exhibit. As far as the brain is concerned, no one incoming signal is the "primary signal" with the others being "context". Each time slice of perceptual experience is a rich tapestry of multi-modal environmental inputs, all of which the brain integrates and incorporates simultaneously. Our results suggest that, across the domains of language and vision, it is surprisingly good at doing that job immediately and continuously.

Acknowledgments

We thank Michael Tanenhaus, Kathy Eberhard, and Julie Sedivy for helpful discussions, and Quinn Hamilton for assistance with data collection. Supported by a Sloan Fellowship in Neuroscience (M.J.S.).

References

Allopenna, P. D., Magnuson, J., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye-movements: evidence for continuous mapping models. *Journal of Memory and Language,* 38, 419-439.

Altmann, G. T. M. & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73, 247-264.

Bridgeman, B. & Aiken, W. (1994). Attentional "popout" and parallel search are separate phenomena. *Investigative Ophthalmology & Visual Science,* **35**, 1623.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge, MA: MIT Press.

Corbetta, M., Shulman, G., Miezin, F., & Petersen, S. (1995). Superior parietal cortex activation during spatial attention shifts and visual feature conjunction. *Science,* 270, 802-805.

Duncan, J. & Humphreys, G. (1989). Visual search and stimulus similarity. *Psychological Review*, 96, 433-458.

Eberhard, K. M., Spivey-Knowlton, M., Sedivy, J. & Tanenhaus, M. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research,* 24, 409.

Eckstein, M. P. (1998). The lower visual search efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological Science* 9, 111.

Egeth, H. E., Virzi, R., & Garbart, H. (1984). Searching for conjunctively defined targets. *Journal of Experimental Psychology: Human Perception and Performance,* 10, 32.

Fodor, J. A. (1983). *Modularity of Mind*. Cambridge, MA: MIT Press.

Friedman-Hill, S. & Wolfe, J. (1995). Second-order parallel processing: Visual search for the odd item in a subset. *Journal of Experimental Psychology: Human Perception and Performance,* 21, 531.

Livingstone, M. & Hubel, D. (1988). Segregation of form, color, movement, and depth: Anatomy, physiology, and perception. *Science,* 240, 740-749.

Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature*, 244, 522-523.

Marslen-Wilson, W. (1975). Sentence perception as an interactive parallel process. *Science*, 189, 226-228.

McGurk, & MacDonald (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.

McLeod, P., Driver, J., & Crisp, J. (1988). Visual search for conjunctions of movement in visual search. *Nature,* 332, 154-155.

Motter, B. C. & Holsapple, J. W. (2000). Cortical image density determines the probability of target discovery during active search. *Vision Research,* 40, 1311 1322.

Motter, B. C. (1993). Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *Journal of Neurophysiology,* 70, 909-919.

Nakayama, K. & Silverman, G. H. (1986). Serial and parallel processing of visual feature conjunctions. *Nature,* 320, 264 265.

Nakayama, K. & Joseph, J. (1998). Attention, pattern recognition, and pop-out in visual search. In R. Parasuraman (Ed.), *The Attentive Brain*. Cambridge, MA: MIT Press. pp. 279-298.

Olds, E. S., Cowan, W. & Jolicoeur, P. (2000). The time-course of pop-out search. *Vision Research,* 40, 891-912.

Palmer, J., Verghese, P., & Pavel, M. (2000). The psychophysics of visual search. *Vision Research,* 40, 1227-1268.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case of impenetrability of visual perception. *Behavioral and Brain Sciences,* 22, 341-423.

Spivey, M. J. & Marian, V. (1999). Cross talk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science,* 10, 281-284.

Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M., & Sedivy, J. C. (in press a). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*.

Spivey, M. J., Tyler, M. J., Eberhard, K. M., & Tanenhaus, M. K. (in press b). Linguistically mediated visual search. *Psychological Science*.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information during spoken language comprehension. *Science*, 268, 1632-1634.

Treisman, A. & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, 12, 97-136.

Wolfe, J. M. (1994). Guided Search 2.0: A revised mode of visual search. *Psychonomic Bulletin & Review,* 1, 202-238.

Wolfe, J. M. (1998). What can 1 million trials tell us about visual search? *Psychological Science,* 9, 33-39.

Zeki, S. (1993). *A Vision of the Brain*. Oxford: Blackwell Scientific (1993).

"Two" Many Optimalities

Òscar Vilarroya (24678ovo@comb.es)
Centre de Recerca en Ciència Cognitiva
Gran de Gràcia, 127, 2-1; 08012 Barcelona, Spain

Abstract

In evolutionary biology a trait is said to be optimal if it maximizes the fitness of the organism, that is, if the trait allows the organism to survive and reproduce better than any other competing trait. In engineering, a design is said to be optimal if it complies with its functional requirements as best as possible. Cognitive science is both a biological and engineering discipline and hence it uses both notions of optimality. Unfortunately, the lack of a clear methodological stance on this issue has made it common for researchers to conflate these two kinds of optimality. In this paper I argue that a strict distinction must be kept in order to avoid inaccurate assumptions.

Cognitive Explanations

Contemporary biological explanations are teleonomic explanations. Teleonomic explanations are those explanations that treat biological traits as adaptations. Adaptations are innovations that make a difference between alternative biological designs embodied in different individuals. Such designs are "chosen" from among alternative designs on the basis of how well they perform in a given environment.

The phenotype of an organism is therefore understood as a collection of functionalities that were added and maintained (i.e., copied over generations) in the design features of a given species *because* these functionalities had the consequence of solving problems that promoted, in some way or other, survival and reproduction (Millikan, 1984). A complete teleonomic explanation would then consist of a step-by-step process in which the scientist must:

(A) Teleonomic Explanation

1. Identify the trait that is likely to be under selection.
2. Identify the adaptive problem that the trait is supposed to solve.
3. Show:

(a) that the trait is specialized for solving the adaptive problem;
(b) that it is unlikely to have arisen by chance alone; and
(c) that it is not better explained as the by-product of mechanisms designed to solve some alternative adaptive problem.

4. Establish the fitness value of the trait in the population.[1]

This sort of story could be the general strategy for *any* discipline which required teleonomic explanations. However, cognitive science has certain particularities that change this methodology. In explaining the cognitive mechanisms of biological organisms, the cognitive scientist may attempt to identify the adaptive problem that the brain is supposed to solve. In reality, however, this turns out to be quite difficult, because the trait—that is, the specialized device within the brain that is responsible for solving the adaptive problem—is not as self-evident as, say, an eye, a liver or a wing. Specialized cognitive mechanisms lie within the circuitry of the brain in such a way that makes them not as obvious as one would like (Barkow, Cosmides & Tooby, 1992).

In order to overcome this problem, cognitive scientists usually invert the first and second steps of the algorithm above:

(B) Cognitive Explanation

1. Identify the adaptive problem that the organism is supposed to solve.
2. Presuppose the trait that is likely to be under selection.
3. Show:

(a) that the trait is specialized for solving the adaptive problem;
(b) that it is unlikely to have arisen by chance alone; and
(c) that it is not better explained as the by-product of mechanisms designed to solve some alternative adaptive problem.

4. Establish the fitness value of the trait in the population.

Inverting the two first steps of algorithm (A) cannot be without consequences. The most obvious is the fact that identifying the function of a trait, such as a wing

[1] In other words, the theorist must establish that the distribution of the trait in the population contributes to the evolutionary notion of fitness, which for present purposes means simply the capacity to survive and reproduce.

or an eye, is much easier when the trait has been identified rather than presupposing the trait when we only know its function. In the first case we can simply observe the trait at work or determine if it is rarely used, etc., or perhaps we might test it under all imaginable circumstances just to see what it does. Cognitive scientists face a much more difficult enterprise, however, since we have to "imagine" the design features of the trait. This is where the reverse-engineering strategy comes in.

Optimality in Reverse-Engineering

Dennett (1995) is perhaps one of the most adamant about defending the reverse engineering strategy in cognitive science. This strategy can be defined as the interpretation of an already existing intelligent artifact or system through an analysis of the design considerations that must have governed its creation. Logically, this overidealizes the design problem, because it presupposes that the trait is *optimally* executed by the cognitive machinery. Thus, reverse-engineering takes cognitive systems to be systems that are designed to solve the problem identified by the theorist; otherwise, the analysis could not get off the ground. As Dennett observes, if cognitive scientists cannot assume that there is a good rationale for the features they observe in cognizers, they cannot even begin their task. Optimality must be the default assumption in cognitive explanations.

A standard way of advancing the reverse-engineering strategy is to resort to Cummins' notion (1983) of functional analysis. Basically, a functional analysis amounts to:

1. System S performs F
2. F can be broken down into $f_1, f_2, f_3 \ldots f_n$
3. S implements $f_1, f_2, f_3 \ldots f_n$

The first step is therefore to establish F, that is, what the system does. The usual way in which one characterizes F is to call upon the computational theory proposed by Marr (1982). In this framework, the theorist must provide an abstract formulation of the information-processing task that defines a given cognitive ability. Peacocke (1986), for example, describes such formulations as characterizations of the information state that the system draws upon.

Now, the notion of "information drawn upon" can be spelled out as follows:

> A state draws upon some information whenever such state carries the information which is causally influential in the operation of the algorithm or mechanism. (Peacocke, 1989, p.102).

Given this definition, explanations based on Peacocke's notion can be seen as a fully causal. For instance, facts about the meaning, syntactic structure, and phonetic form of linguistic expressions are causally explained by facts about the information drawn upon by algorithms or mechanisms in the language-user (ibid., p.113). Thus, if a system draws upon the correct information, then the explanation of system is correct regardless of the detailed algorithm that the system uses.

Be this as it may, the computational characterization of a problem determines the specifications by which the reverse-engineering must proceed. These specifications are taken at face value and are considered sufficient to establish the design features of the mechanism. Cummins (1983) further argues that such characterizations should proceed by decomposing them into a set of simpler capacities that are to be explained by subsumption. The overall capacity is thus explained in terms of the contributing capacities of parts of the system, and the function of a given item is its contribution to the overall capacity. Such design theories of functions define the function of a mechanism or process in terms of the roles they might play, that is, in terms of their contribution to some capacity of the system to which the process or mechanism belongs. In short, design theories relativize functions to capacities of containing systems.

Optimality in Evolutionary Biology

The previous sense of optimality must be distinguished from the notion of optimality that is used in evolutionary biology. To begin with, the biological notion of an optimum does not imply an optimal design, as it does in engineering. Rather, it refers to a solution to a given adaptive problem that maximizes the fitness of the organism in the adaptive situation.

A biological optimum can be said to be the point at which the difference between costs and benefits of environmental and genetic variables (e.g., amount of food, energy requirements, distribution of the trait, alternative phenotypes, etc.) is maximized (see Figure 1). Thus, the role of a trait as an adaptation must be established by considering the manner in which such a trait contributes to the optimum. This makes the notion of optimality in biology and engineering orthogonal to one another:

Biological optimality: Natural selection favors the trait that maximizes the organism's *fitness*.

Reverse-engineering optimality: A mechanism is *designed* to comply with its function.

The fact of the matter is that a trait need not be optimally designed to be adaptive: what is optimal is the fitness value of the trait, not its design characteristics. Evolutionary solutions must, on this view, only be "selectively efficient," that is, they need only to comply with adaptive requirements. It follows, then, that the notion of optimal design should be detached from the notion of fitness value.

That natural selection is only susceptible to the fitness-value maximizations is anything but surprising: evolution by natural selection only requires that biological systems be minimally effective (stay alive and leave offspring) with respect to their of adaptive problems. Hence, rather than actively designing and building organisms that are well-adapted to the world, nature eliminates

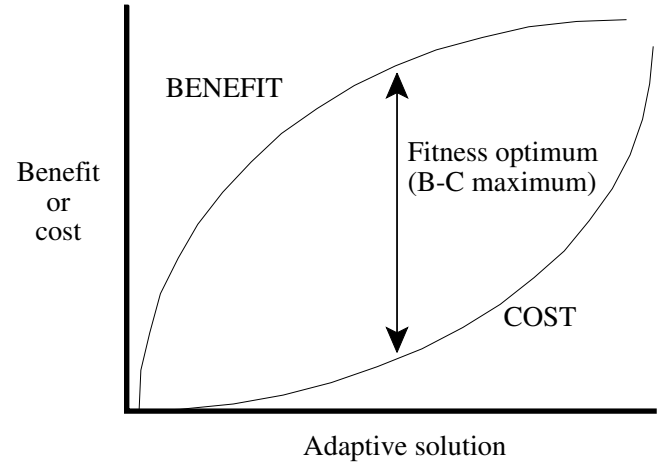


Figure 1: The optimum for a given adaptive solution is the point in which the benefit/cost relationship is maximized.

those that are too ill-suited for survival and reproduction. In other words, biological systems are simply not designed by engineers. Design and evolution are different precisely because they have different strategies open to them. An engineer may build a system out of an analysis of the problem, and thus may go from the problem to the solution. This is not a possibility for biological systems: biological systems are blind to the solution until they have stumbled upon it.

The bottom line, then, is that we should explain how cognitive systems are selected for and maintained by taking into account not only the adaptive problem itself, but also their resources and the environment in which they are evolving.

Conflating the Two Optimalities

The previous discussion makes it advisable to maintain the engineering and the biological optimalities separate. However, some cognitive scientists (e.g., Barkow, Cosmides & Tooby, 1992) seem to conflate both notions. According to these theorists, what must guide the design specifications of cognitive mechanisms is a computational characterization (Marr, 1982) of the adaptive problems that these mechanisms were meant to solve. These specifications are considered sufficient to establish the design features of the mechanism:

> In effect, knowledge of the adaptive problems humans faced, described in explicitly computational terms, can function as a kind of Rosetta Stone: It allows the bewildering array of contents effects that cognitive psychologists routinely encounter—and usually disregard—to be translated into meaningful statements about the structure of the mind. (Cosmides & Tooby, 1992, p.221)

In other words, these theorists take the brain to be the seat of specialized mechanisms that are optimally designed, in an engineering sense, to solve specific adaptive problems.

It is of course possible for engineering and biological characterizations to coincide for a given trait and a given organism. Nonetheless, counterexamples abound (e.g., Ullman, 1996; Steels, 1994; Dehaene, 1997; Cooper & Munger, 1993;). Consider the well-known example of the sight-strike-feed mechanism of the frog (Gilman, 1996). Frogs catch flies by way of a strike with their tongue. It is assumed that mediating between the environmental presence of a fly and the motor response of the tongue strike there is some sort of mechanism that registers the fly's presence in the vicinity of the frog. That is, the presence of the fly causes the relevant mechanism to go into state *S*, and its being in state *S* causes the tongue to strike.

This story goes on to assume that the information drawn upon by state *S* is that of "fly", "fly, there," or "edible bug, there," since this information can be derived from that the fact that the function of the frog's sight-strike-feed mechanism mechanism is to detect the presence of flies. Yet, an analysis of the frog's cognitive system indicates that the best account of the system's function is in fact detecting "little ambient black things." Specifically, the function of the mechanism is to mediate between little ambient black things and the frog's tongue strike.

This means that the frog's mechanism is functioning optimally even when the frog strikes at a little ambient black thing that is not a fly but a BB-gun pellet that happens to be in the vicinity. To be sure, from a reverse-engineering point of view, the system is not optimally designed to catch flies. However, from a biological point of view, it might be the optimal system. The reason is quite simple. The cost of "fly-detecting" mechanism may outweigh the cost of eating, say, lead pellets. The guarantee that a frog with such an less-than-perfect mechanism could have survived and reproduced is provided

by the contingent fact that, during natural selection, a sufficient number of little ambient black things in the frog's environment were flies (or edible bugs). The combination of the benefits (which should include adequate feeding) with the costs (which should include design-building costs) shows that a better design need not mean better fitness, which is what would be predicted if only design were considered. Accordingly, in some situations a better design can actually mean a drop in fitness.

It might be objected that the fact that frogs also flick their tongues out at little black things that happen not to be flies is an empirical discovery and, hence, either sense of optimality could have been wrong about what frogs would do when confronted with BB-gun pellets. For example, an evolutionary biologist might predict that natural selection would favor the trait if the trait were specifically tailored to fly catching, since this would maximize fitness. Yet this prediction would have been wrong. A reverse-engineering perspective, by contrast, might well have made the correct prediction: striking at little black things that are not flies might be seen as an acceptable amount of noise, and not necessarily an unoptimally designed fly-catching mechanism. Such being the case, the problem, it might be argued, does not really have anything to do with the two different notions of optimality, but rather with the claims one is making about a particular trait and what sort of evidence should be used in evaluating those claims.

It seems to me that such an objection would miss the point of the argument, which is to uncover two different methodological strategies, and not competence in hypothesis formation. I will illustrate the problem with another example. Peacocke (1993) has argued that the use of particular kinds of physical principles is constitutive of the capacity of normal mature subjects to reason about and predict object motions. Such a constitutive basis is held to underlie the remarkable precision of our perceptual systems in extracting and using the motion of objects in space. Examples of this capacity include our ability to anticipate the trajectories of objects in order to intercept, follow, or avoid them.

As is well known, two general types of information are used in classical physics to describe the behavior of moving objects. On the one hand, kinematic information describes the pure motion of bodies without regard to mass (i.e., the position, velocity and acceleration of an object). On the other, dynamics describes the forces causing movement or acting on objects with mass. According to Peacocke (ibid.), in order to qualify as being able to reason about objects, we must attribute to humans the capacity to reason according to dynamic principles. This would correspond to what I have described as the task characterization, which is a normative description: it is what the system must do in order for its behavior to be selectively efficient (e.g., avoid falling stones).

If we employ a reverse-engineering strategy, the task characterization of reasoning and predicting object motions (*qua* dynamic computation) will be all that we need to analyze the system that accounts for such a capacity. If, on the other hand, we employ an evolutionary strategy, then we will have to develop a model of adaptation (see, for example, Parker & Maynard Smith, 1990). Such a model will have to consider competing alternatives that exist in an adaptive scenario. For instance, we can assume that we should evaluate the performance of a system that predicts object motion according to kinematic variables, and another according to dynamic variables. This comparison should establish the performance of each system, not in isolation but as a part of the whole organism-environment interaction. Once this is done we will be able to consider the costs of either system, in terms of design and computation.

It is very conceivable that this model might yield an outcome that is very different from the reverse-engineering analysis. It could, for example, provide the hypothesis that the kinematic system is the most adaptive solution because it satisfies the task of predicting object trajectories in a way that outweighs the cost of a much more complex, yet more optimally designed, computational system that computes dynamic variables. Among other things, the errors induced by a kinematic system may not be unacceptably gross and may be easily compensated by the continuous activation of the perceptual system. This would be congruent with empirical research such as Cooper and Munger (1993).

The point, then, is this: if we had relied the reverse-engineering strategy we would not have reached the correct analysis. This is not because we would have assumed an incorrect claim but because we simply would have employed the wrong optimality strategy.

Having said this, it is no doubt true that the distinction between the engineering and biological optimalities might not be an easy matter, at least not at first blush. On the one hand, the functionality of the system (e.g., detecting flies) is amenable to both reverse-engineering and ecological analyses; on the other, it is not always clear how to establish the parameters of the fitness-maximization process that constrains adaptive cognitive traits. The latter might not be impossible to establish in cognitive science (Vilarroya, 2001). The former requires changes in algorithm (B).

Cognitive Explanation Revisited

In my opinion the explanatory strategy of cognitive science cannot be simply an inversion of the first steps of the teleonomic explanation. It is not enough to identify the adaptive problem and then infer the mechanism. Rather, we need to complement the assumption about a trait's design with a characterization of how the adaptation might have appeared over evolutionary time. Fortunately, we have the elements to proceed to the different steps necessary to complete a cognitive explanation. In order to do that, we should divide the first step of algorithm (B) into two substeps, namely:

(B′) Cognitive Explanation

1. Characterize:

(a) the adaptive problem that the organism is supposed to solve; and
(b) the fitness-maximization process.

2. Presuppose the trait that is likely to be under selection.
3. Show:

(a) that the trait is specialized for solving the adaptive problem;
(b) that it is unlikely to have arisen by chance alone; and
(c) that it is not better explained as the by-product of mechanisms designed to solve some alternative adaptive problem.

4. Establish the fitness value of the trait in the population.

The explanation should proceed as follows. The first sub-step should yield the informational-theoretic characterization (Marr, 1982) of the functional requirements needed to satisfy the adaptive problem. Specifically, we need to indicate the requirement imposed by the adaptive problem that the system should satisfy in an idealized situation.

Once this characterization has been established, then the theorist must proceed to characterize the fitness-maximization process (including the adaptive requirements that are to be satisfied by the organism in such a process). Then, the theorist should verify whether the computational characterization of the adaptive problem is compatible with the optimum established in the fitness-maximization process. If both draw upon the same information, then the characterization of the adaptive problem can be used in conjunction with reverse-engineering methodology. This should yield the assumed design specifications for the trait. If, on the other hand, the the adaptive problem's computational characterization is not compatible with the fitness-maximization optimum, then the functional requirements of the fitness-maximization process should guide the design assumptions. As the case of the frog has shown, the fitness-maximization account allowed the assumption that the trait should be designed to detect "little ambient black things," rather the one offered by the adaptive problem which would have been "fly-there."

The characterization of the fitness-maximization process in cognitive science is, unfortunately, not a straightforward operation, as I have shown elsewhere (Vilarroya, in press). It is actually a complex process because, among other things, there is an essentially open-ended set of factors that influence just where the cost-benefit curve reaches its maximum. What allows an individual, or a group of individuals, to survive and leave offspring depends precisely on their biological constitution and the exact characteristics of the surrounding environment with which they interact.

Nonetheless, the elements in this characterization are objective. Therefore one can hope to make them explicit, and thus provide an adaptive characterization of the trade-off between the costs as well as the benefits of available solutions. This will (or should) eventually yield a description of the functional account of the cognitive system, if the analyst takes into account: (a) the nature of the adaptive problem itself, (b) the analysis of the system's resources, (c) the environment and interaction with competitors, as well as (d) the way in which all these elements interact.

How can we apply this characterization in the case of the frog? I believe that there is a way to account for the paradox that the cognitive system of the frog accords with a characterization of the adaptive problem, even though it is not the characterization of the mechanisms that accounts for the adaptive capacity. For one thing, we already have an account of the adaptive problem that the system has to solve: identify flies. This is a normative description; it is what the system must do in order for its behavior to be selectively efficient. For another, we have the functional requirements that the fitness-maximization process imposes: identify "little ambient black things." The evolutionary rationale behind this solution is consistent with the assumption that the extra computational cost of taking only flies into account (over and above other small dark ambient things) arguably outweighs the small increase in accuracy that would be gained from doing so.

In sum, a system may seem to accord with a certain functionality (e.g., identifying flies) that is, in actuality, different from the description of the structure of the cognitive system (e.g., identifying "little ambient black things"). Accordingly, such a system may in fact take advantage of mechanisms that are not specifically designed to deal the problem at hand. This does not mean that the solution is somehow deficient. The fact that the fly-catching mechanism of the frog is not sensitive only to flies does not mean that it cannot identify flies. It can and it does.

Conclusion

Cognitive science uses two distinct notions of optimality: engineering optimality and biological optimality. In engineering, optimality refers to design ideals whereas, in biology, optimality refers to fitness maximization. While conflation of the two concepts is understandable, neglecting the distinction entails incurring risk of arriving at a mistaken conclusion. Fitness value and designs are therefore best analyzed separately.

Acknowledgments

I would like to thank the following friends and colleagues for their help and advice: Antoni Gomila, David Casacuberta, Joseph Hilferty, Joan Carles Soliva, Javier Valenzuela, Maria Verdaguer, and Agustín Vicente.

References

Barkow, J., Cosmides, L. & Tooby, J., (Eds.) (1992). *The Adapted Mind: Evolutionary Psychology and the*

Generation of Culture. New York: Oxford University Press.

Cooper, L.A. & Munger, M.P. (1993). Extrapolating and Remembering Positions along Cognitive Trajectories: Uses and Limitations of Analogies to Physical Motion. In Eilan, N., R. McCarthy & Brewer, B. (Eds.), *Spatial Representation*. Oxford: Basil Blackwell.

Cosmides, L. & Tooby, J. (1992). Cognitive Adaptations for Social Exchange. In Barkow, J., L. Cosmides, & Tooby, J. (Eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. New York: Oxford University Press.

Cummins, R. (1983). *The Nature of Psychological Explanation*. Cambridge: MIT Press.

Dehaene, S. (1997). *The Number Sense: How the Mind Creates Mathematics*. New York: Oxford University Press.

Dennett, D.C. (1995). Cognitive Science as Reverse Engineering: Several Meanings of "Top Down" and "Bottom Up." In Prawitz, D., Skyrms, B. & Westerstahl, D. (Eds.), *Proceedings of the 9th International Congress of Logic, Methodology and Philosophy of Science*. Dordecht: North Holland.

Franks, B. (1995). On Explanation in the Cognitive Sciences: Competence, Idealization, and the Failure of the Classical Cascade. *British Journal of Philosophy of Science 46*, 475-502.

Gilman, D. (1996). Optimization and Simplicity: Computational Vision and Biological Explanation. *Synthese*, *107*, 293-323.

Marr, D. (1982). *Vision*. Cambridge, MA: MIT Press.

Millikan, R. (1984). *Language, Thought and Other Biological Categories*. Cambridge, MA: MIT Press.

Parker, GA, & Maynard Smith. (1990). Optimality theory in evolutionary biology. *Nature* 348, 27-33.

Peacocke, C. (1986). Explanation in Computational Psychology: Language, Perception and Level 1.5. *Mind and Language 1*, 101-123.

Peacocke, C. (1989). When is a Grammar Psychologically Real? In Alexander, G. (Ed.), *Reflections on Chomsky*. Oxford: Basil Blackwell.

Peacocke, C. (1993). Intuitive Mechanics, Psychological Reality and the Idea of a Material Object. In Eilan, N., McCarthy, R. & Brewer, B. (Eds.), *Spatial Representation*. Oxford: Basil Blackwell.

Steels, L. (1994). The Artificial Life Roots of Artificial Intelligence. *Artificial Life*, *1*, 75-110.

Ullman, S. (1996). *High-Level Vision*. Cambridge, MA: MIT Press.

Vilarroya, Ò. (2001). From Functional "Mess" to Bounded Functionality. *Minds and Machines* 11, *in press*.

Generalization In Simple Recurrent Networks

Marius Vilcu (mvilcu@cs.sfu.ca)
School of Computing Science
Simon Fraser University, 8888 University Drive, Burnaby, Canada, V5A 1S6

Robert F. Hadley (hadley@cs.sfu.ca)
School of Computing Science
Simon Fraser University, 8888 University Drive, Burnaby, Canada, V5A 1S6

Abstract

In this paper we examine Elman's position (1999) on generalization in simple recurrent networks. Elman's simulation is a response to Marcus et al.'s (1999) experiment with infants; specifically their ability to differentiate between novel sequences of syllables of the form ABA and ABB. Elman contends that SRNs can learn to generalize to novel stimuli, just as Marcus et al's infants did. However, we believe that Elman's conclusions are overstated. Specifically, we performed large batch experiments involving simple recurrent networks with differing data sets. Our results showed that SRNs are much less successful than Elman asserted, although there is a weak tendency for networks to respond meaningfully, rather than randomly, to input stimuli.

Introduction

In a recent paper, Elman (1999) casts doubt upon the widely noted results of Marcus et al. (1999). In the Marcus et al.'s experiment, 7-month old infants were habituated to sequences of syllables of the form ABA or ABB (e.g., "we di we" or "le di di"). Marcus et al. found that infants showed an attentional preference for novel test sequences of syllables (which we call "sentences"), which differed from the habituation stimuli[1]. Marcus et al. argue that the reason for this behavior is the fact that infants extracted "algebra-like rules that represent relationships between placeholders (variables)" (1999). They also concluded that simple recurrent networks (and, in general, all networks whose training is based on backpropagation of error) were not able to display this kind of behavior because they could not generalize outside the training space.

The issue of generalization outside the training space was previously addressed in Niklasson and van Gelder (1994), and Marcus (1998). In essence, the training space represents the n-dimensional hyperplane delimited by the set of training vectors. We say that a connectionist model generalizes to novel stimuli when correct output is reliably produced for an input item that was not included in the training set (i.e., the network was never trained on that stimulus in any position within its input layer). Marcus maintains that a neural network trained with the backpropagation algorithm (or any variant of it) is not able to display such a behavior, because the innate structure of the backpropagation algorithm[2] precludes the network from generalizing to nodes that have not been specifically trained.

Elman agrees that the Marcus experiment does "indicate that infants discriminated the difference between the two types of sequences" (1999), but he believes that this result may be explained by the relationship between the last two syllables: infants were able to distinguish that in one case the last two syllables were identical (ABB), and in the other case the last two syllables were different (ABA). Moreover, Elman maintains that it is feasible for a simple recurrent network to perform this same task, provided the network is presented with the same background knowledge as infants have (in particular, an exposure to a wide range of syllables that infants have before participating in the experiment).

Having said that, Elman performs an experiment involving an SRN that aims to simulate the Marcus et al.'s experiment. There are three phases in Elman's simulation: 1) the pre-training period, corresponding to the prior experience of the infants in learning to recognize syllables; 2) a second phase corresponding to the habituation task that infants encountered (presenting ABA and ABB sentences); 3) a testing phase involving novel stimuli, as in the infants' experiment. At the end of his simulation, Elman concludes that his results "clearly indicate that the network learned to extend the ABA vs. ABB generalization to novel stimuli" (1999).

Granting Elman's basic assumptions, we constructed an experiment that mimics his simulation. We did not

[1] For example, after habituated to ABA sequences, the infants spent more time recognizing novel test sequences of the form ABB than did for ABA sequences, and vice versa.

[2] The weights connecting a given output node are trained independently of the weights connecting any other output node. Consequently, the set of weights connecting one output unit to its input units is entirely independent of the set of weights feeding all other output units. This is called input-output independence, and it is believed to be the major weak point of backpropagation neural networks. It is less clear that the problem arises for competitive learning networks, however. See Hadley et al (1998) for details.

have access to all Elman's data, but we used the same Plunkett and Marchman's (1993) distinctive feature notation of consonants and vowels that Elman employed in his experiment. However, since the results we obtained led us to a different conclusion than Elman's, in order to have a more complete picture of the performances of simple recurrent networks, we created a variety of data sets by changing the degree of overlapping units in the training/testing vectors. Also, to be sure that the results were not obtained by chance, we performed batch training, i.e., at least 64 different training-test sessions were carried out for each individual training corpus (i.e., 64 or 128 different weight initializations were assigned to the basic configuration, resulting in 64 or 128 separate networks trained on each data set).

Basic structure of Elman's and our experiments

A simple recurrent network architecture was used for all experiments. The input layer contains 12 or 24 units (depending on the experiment; see details below). The number of hidden/context units was varied between 10 and 40. The output layer contains two units; one was used only during the pre-training phase, while the other unit was only used during the sentence habituation and testing phases.

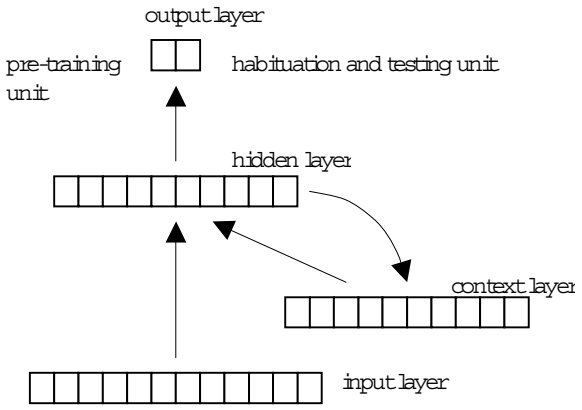


Fig. 1. Architecture of the network

The data set deployed in the pre-training phase contained 50,000 syllables (separate tokens) from the full set of 120 possible types. Each syllable was presented to the network, one at a time, and the SRN was trained to distinguish between the current syllable and the previous one (whether or not they are identical). Only one of the two output units was used during this supervised training.

The habituation phase followed the pre-training phase. During this phase the same network was presented with 32 distinct sentences formed with 8 different syllables (these 8 syllables also occurred in the set of 120 types of syllables employed in the pre-training set). The 32 sentences were generated from the ABA and ABB grammars (16 ABA sentences, and 16 ABB sentences). Each sentence was presented to the network, one syllable at a time. Following the last syllable of a sentence, the network was trained to output a 0 in the case of ABA sentences, and a 1 in the case of ABB sentences. During this training phase only the second output unit was used (the one not used during pre-training). Interestingly, the weights were modified only after the last syllable of the current sentence was presented (no training occurred following the first two syllables). This was done in order to ensure that the network would learn to make discriminations the same way as the infants presumably would, using similar stimuli.

For testing, four sentences were used, formed with 4 "relevantly novel" syllables (i.e., these syllables appeared in the 120-syllable pre-training set, but not in the training corpus). Two sentences had the form ABA, and the other two had the form ABB. Again, the second output unit was used to monitor the network's responses.

Before presenting our results, we would like to clarify the following issues:

1) Because we were not able to have exactly the data set that Elman used, we generated our patterns based on Plunkett & Marchman's (1993) feature representation of consonants and vowels. Since Elman encoded his stimuli using the same notation, we believe that the difference between our data set and Elman's is minimal and arguably insignificant to the outcome of the experiment.
2) Elman's main objective was to challenge Marcus' assertion that SRNs are not able to generalize outside the training space. However, we believe his claims are overstated. Although a minority of sessions in our batch jobs was as good as Elman's, in general, we found that the SRN did not perform as well as Elman maintains.

As noted above, all experiments were based on Elman's simulation. Between our experiments and Elman's there were a few differences, however. These consisted in the way in which the data sets were created and the way the results were computed. Our first data set very closely resembles Elman's representation of vectors, both corpora being based on the same distributed representation of syllables. Since the results we obtained for this first data set offered only little evidence to support Elman's position, we have created a second corpus of patterns, by changing Elman's original vectors in order to have a more uniform and more overlapping data set (see below the description of Experiment 2). Lastly, our third data set employs completely non-overlapping vectors, i.e. we used a localist representation of input patterns.

Experiment 1

The input corpus for this experiment was very close to the Elman's data set. We used distributed representations to create the patterns: each syllable had 12 phonetic features, each syllable being made up from a consonant followed by a vowel. All the syllables were generated randomly using the whole set of letters, and the patterns were created based on Plunkett & Marchman's (1993) notation of each letter[3]. We created 120 vectors this way. All of these patterns were used in the pre-training phase, while 8 of them were employed during training and other 4 vectors were used for testing.

For example, here are 2 of the 8 training syllables and 2 of 4 testing syllables:

training

```
mo -1  1 -1 -1  1  1 -1  1 -1  1 -1  1
wu -1  1 -1  1  1  1  1  1  1  1 -1  1
```

testing

```
za -1  1  1 -1 -1  1  1  1  1  1 -1 -1
fe -1 -1  1 -1  1  1  1  1 -1 -1  1 -1
```

We tried to generate as diverse and random data set as possible, like infants are presumably exposed to prior to participating in the Marcus et al's experiment. However, our results showed that these patterns were not very "friendly" to our SRNs, and the networks were not nearly as successful as infants in discriminating those sentences.

One of the characteristics of this data set was that, because of the randomness of patterns, many of the testing vectors were very different from the vectors employed in training. For instance, the average distance[4] among training vectors was about 3-4 bits, while the difference between training and testing patterns exceeded 6-7 bits. In our opinion, this contrast among patterns is responsible for making the testing session difficult.

Experiment 2

Since the results based on the first data set failed to prove Elman's strong claims, we generated a different corpus of stimuli, trying to make the training process successful. Consequently, we have manually created 12 vectors (8 vectors are used in training, the other 4 in testing). The remainder of 108 vectors have been generated randomly. All these patterns have been distributed uniformly between the two sets of stimuli (training and testing). In this way, the distance among vectors within the same set of stimuli (whether training or testing) was similar to the distance between vectors found in the different corpora.

For example, here are a few of the vectors used in this experiment:

training

```
-1 -1 -1  1  1 -1 -1  1 -1  1  1  1
 1 -1 -1  1  1 -1 -1  1  1 -1 -1  1
```

testing

```
 1 -1  1 -1  1 -1  1 -1  1 -1  1 -1
-1 -1  1  1 -1 -1  1  1 -1 -1  1  1
```

In this case, the average difference among the vectors within the same corpora is about 4-6 bits. This value is close to the difference between training stimuli and testing stimuli (6-7 bits). Because the patterns are uniformly spread across the training and testing sets, this represented a training advantage for networks. However, this tactic further reduces the novelty of the test "sentences".

Experiment 3

The third experiment involved a rather different data set. This employed completely non-overlapping vectors. As a result, the vectors were larger (24 bits, instead of 12).

For example, here are 4 of 12 training/testing vectors (the rest of the 108 vectors used during pre-training were generated randomly):

```
1) 1 -1  1 -1 -1 -1 -1 -1 -1 -1 -1 -1
-1  1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
2)-1  1 -1  1 -1 -1 -1 -1 -1 -1 -1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
3)-1 -1 -1 -1  1 -1  1 -1 -1 -1 -1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
4)-1 -1 -1 -1 -1  1 -1  1 -1 -1 -1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
```

As it may be seen, the patterns were generated by moving a 4-bit frame [1 -1 1 -1] along the 24 bit vectors. In this way, the resulting vectors do not overlap, and the distance among all vectors is the same (4 bits).

Although this corpus of stimuli had little in common with Elman's data, we wanted to examine the performance of SRNs in the case of non-overlapping vectors and to address Marcus' issue about generalization to genuinely novel items.

Results

The performances of SRNs (and, in general, any network using the backpropagation algorithm) are influenced by several training parameters, such as initial weights, learning rate, etc. Usually, the initialization of weights is performed randomly and if training parameters are not chosen properly (especially the learning rate), the network may end up in a region of local minimum of the error function. One way to

[3] For example, the pattern for syllable "da" was a 12-bit vector created by concatenating the 6-bit feature representation of "d" with the 6-bit feature representation of "a".

[4] The distance between two vectors is given by the number of bits by which the two vectors differ.

reduce this liability is to perform a batch experiment, to test the network with a large number of different weight initializations and training parameters. Another advantage of this approach is that at the end of the batch sessions we will have a more precise picture of the behavior of the networks, and also know whether or not the results are generated accidentally.

Significantly, in a series of preliminary experiments, we found that, very often, the weight initializations we used determined poor results for networks, regardless of training parameters (including hidden layer size). Therefore, we decided to perform at least 64 different training-test sessions for each of our 3 experimental designs, each session using a different weight initialization (there were 64 sessions for the first two experiments, and 128 different sessions for the third experiment). In this way, we generated at least 64 separate trained networks for each batch experiment[5].

We chose to use two criteria in order to evaluate the results:

(1) Our first criterion was simply based on the percentage of "acceptable" results. We say that a network generates acceptable results when it grammatically categorizes each presented test sentence within 30% of its target value. Since there are 4 test sentences (2 ABA sentences, and 2 ABB sentences), we have 4 output values to record for each of the 64 networks (let's say, A, B, C, D are the network outputs for the 4 test sentences). Having 0 as the target value for A and B, and 1 as the target value for C and D, an acceptable result is an output value less than .3 for A and B, and greater than .7 for C and D.

(2) Although the above-mentioned criterion is very tolerant, Elman's results would not be counted as acceptable in conformity with this criterion (one of Elman's responses is about .6, outside the 30% error margin; see below for more details). However, not to reject Elman's approach on a-priori grounds, we adopted a second, more forgiving criterion: we consider a result as "acceptable" if only 3 of the 4 responses are within 30% of their target values, while the remaining response is within 45% of its target value.

Since our extensive set of training-test results were significantly different from Elman's isolated result, we tried to see whether, at least, they were better than mere chance. For the first criterion, the chance value is given by the probability that all 4 test sentences are fortuitously, correctly classified, i.e., the network outputs are within 30% of the target values. Clearly, the chance probability that the network outputs a value in the target range, for any of the 4 test sentences, is .3. Therefore, the probability that all 4 sentences are correctly recognized is .0081 (= .3 x .3 x .3 x .3), or .81%. This represents the "chance" value, which we compared all our results to. For the second criterion, the chance value is slightly different. Here, in order to correctly categorize purely by chance, networks should report output values within 30% of the target values for 3 sentences, and within 45% for the 4th sentence. Consequently, the probability for that happening is .01215 (= .3 x .3 x .3 x .45), or 1.215%.

[5] One could metaphorically regard these trained networks as the infants involved in the Marcus et al.'s experiment.

Experiment 1

This experiment is closest to Elman's simulation. Results reported by Elman (1999) were as follows:

	response	target		response	target
A	0.004	0	C	0.853	1
B	0.008	0	D	0.622	1

Even though the network's response for the last sentence was very close to the chance value of .5, Elman asserted, "these responses clearly indicate that the network learned to extend the ABA vs. ABB generalization to novel stimuli" (1999). In our view, based only on these results, Elman overstates the facts. In accord with our first evaluation criterion, his result would not even have been considered acceptable. We devised the second criterion, even more lenient than the first one, in order to cover Elman's result. In any case, in our extended series of experiments, we found that the responses of our networks were highly dependent on weight initializations.

We performed numerous batch experiments, systematically varying, in all combinations, the available parameters values: learning rate (between 0.01 and 0.1), the number of hidden/context units (between 10 and 40), the momentum (0 and 0.5), weight initialization (within the interval [-1, 1], or [-0.1, 0.1]). The best results were obtained for 30 hidden/context units, a learning rate of 0.01, momentum 0 and weight initialization within [-1, 1].

Specifically, for this first experiment, our results were:

(1) of the 64 trained networks, 15 generated acceptable results in conformity to the first evaluation criterion; thus, the percentage of acceptable results is 15/64 x 100 = 23.43%;

(2) evaluating with the second, more lenient, criterion, the percentage of acceptable results is 23/64 x 100 = 35.93%;

We believe the results lend, at best, weak support to Elmans claims. A percentage of good results around 30% cannot lead us to the conclusion that "the network learned to extend the ABA vs. ABB generalization to novel stimuli", as Elman asserted (1999). Granted, the results are significantly larger than the chance values (.81% for the first criterion, and 1.215% for the second one), which means that there is a tendency for the networks to train in such a fashion that they give meaningful, rather than random results.

As noted earlier, these results might be partially explained by the randomness of patterns used in this

experiment. There were instances when the training vectors were very different from the vectors used for testing (up to 90% of the bits were different).

To prove that a different corpus of stimuli can generate better results, we performed a second set of tests, making the data set more uniform and decreasing the distance between training and testing patterns. Here are the details:

Experiment 2

As noted earlier, most part of the 120 patterns used in this second experiment were created randomly, except the 12 vectors employed in the training (habituation) and testing phases. These 12 vectors were generated manually and distributed uniformly over the training and testing sets in order to have a similar distance among all the vectors.

In this case, the average difference between training and testing vectors is about 6 bits, close to the distance among vectors within the same set (whether training or testing), which is about 7 bits.

We varied many training parameters in this case too, and we obtained the best result for 40 hidden/context units, a learning rate of 0.1 and momentum 0.5.

As expected, the results were substantially better:

(1) there were 40 trained networks (out of 64) whose responses were acceptable in conformity with the first evaluation criterion; thus, the percentage of acceptable results is 40/64 x 100 = 62.5% ;

(2) there were 42 trained networks that responded acceptably in accord with the second criterion; the percentage of acceptable results is: 42/64 = 65.62% ;

Noteworthy, these results were obtained for a number of 40 hidden/context units. When using 10 hidden/context units (as Elman presumably did), the results were worse: 29.68% in accord with the first criterion, and 34.37% evaluating with the second criterion.

Although the percentages of 62.5% or 65.62% of successfully trained networks are not impressive, in contrast with the chance value of .81% (and 1.215% respectively), they represent a significant result (the probability to respond acceptably, in conformity to our criteria, is 80 times greater than the probability by mere chance). Therefore, this experiment demonstrates more convincingly what we noted earlier: there clearly is a tendency for the networks to train in such a fashion that they give meaningful, rather than random results. However, we must bear in mind that the training regime now under consideration does not satisfy the conditions for generalization outside the training space.

Experiment 3

The third experiment differs from the first two with respect to the type of the vectors involved: here we used completely non-overlapping vectors, because we wanted to address Marcus' challenge of generalization outside the training space. Thus, we tried to discover whether simple recurrent networks are indeed able to generalize to novel stimuli.

Initially, it would seem that our testing patterns were not novel to the network (since they also appeared in the pre-training set). But, there are two arguments behind our assumption that the testing vectors are actually novel:

- the output unit used during pre-training is different from the output unit used during habituation (second training). Since the representation of patterns is localist and the training algorithm is backpropagation, these two output units are purportedly independent: the training of one unit should not influence the other unit, as Marcus argued (1998).
- the training regimes used during the pre-training and habituation phases are different (one algorithm teaches the network to determine whether or not consecutive syllables are identical, while the other one teaches the network to differentiate between ABA and ABB sentences). Since the testing vectors do not appear in the training data set used during sentence habituation, they are novel to the network in the relevant sense.

For this experiment we performed two sets of tests, both involving 128 separate training/testing sessions. Although 64 trained networks are presumably enough to form a complete picture of the behavior of networks, we wanted to see whether or not the general tendency noted previously was repeatable for a substantially larger batch experiment. The answer was affirmative.

The first set of experiments employed a test corpus of 4 sentences, exactly the same number of sentences used by Elman, and by us in the experiments 1 and 2. The results were as follows:

(1) there were 8 successfully trained networks (out of 128); thus, the percentage of well-trained networks was, in conformity with the first criterion, 8/128 = 6.25% ;

(2) there were 14 trained networks which responded acceptably in accord with the second criterion; the percentage of acceptable results is: 14/128 = 10.93% ;

Although these values are much less impressive than those of the previous experiment, they still are better than chance. Of course, the absolute percentage of successful networks (6.25, or even 10.93) is small, indicating that SRNs have problems dealing with novel stimuli. However, it is still substantially greater than .81 (or 1.215 for the second criterion), which would have been obtained by pure chance.

However, for the second set of tests we expanded the test corpus to 30 sentences. In this case, none of the 128 trained networks output good results in accord with any of the two criteria. This result was consistent for different training parameters, such as learning rate and number of hidden/context units.

Discussion

In the three experiments described herein, we have systematically varied a wide range of parameters. Indeed, in the case of Experiment 1, were Elman's data set is very closely approximated, we have parametrically varied not only the learning rate and weight initialization range, but also the hidden layer size (which Elman did not do). On the basis of all three experiments described above, we believe it is fair to say that Elman's case has been substantially overstated.

On the other hand, certain of our results may lend some modest confirmation to Elman's position, at least with respect to the very simple syntax employed in the Marcus et al experiment. To be sure, in the case of Experiment 1, which most closely approximated Elman's training data, the percentage of successfully trained networks was only 23.43%. However, this percentage is far above the purely chance values that we have cited. In addition, we have shown that even when all input vectors within a given training corpus are completely non-overlapping (Experiment 3), as many as 6% of trained networks satisfy our "least forgiving" criterion of success, at least when the test corpus contained just 4 sentences (as in Elman's case). Significantly, though, when the test corpus for Experiment 3 was expanded to contain 30 novel sentences, no positive results whatsoever were obtained even when our more lenient "success criterion" was used. This outcome lends clear support to Marcus' claims on "generalization outside the training space" -- at least with respect to the infant learning experiment described by Marcus et al (1999).

Finally, we must also emphasize that, except in Experiment 2 (where we modified the syllable vectors to ensure that training and test input vectors were much more similar), the preponderance of trained networks failed to satisfy even the most forgiving success-criterion adopted here. More importantly, we have replicated the design of Experiment 1 using two modestly more complex grammars, and have obtained only negative results. In particular, when the grammars (ABCA vs. ABCB) and ABCDA vs. ABCDB) were employed, we were unable to train even a single network successfully (from a batch of 64 networks). This strongly suggests that the SRN architecture deployed in Elman's "refutation" of Marcus is incapable of abstracting the underlying structure of anything but the very simplest of grammars. Our view is that the "grammar" deployed by Marcus et al (1999) is perhaps too simple to present a useful challenge to eliminative connectionist networks. A desirable step for future research would be to repeat the "human infant experiment" using the modestly more complex grammars just cited.

References

Elman, J. (1998). Generalization, simple recurrent networks, and the emergence of structure. M A. Gernsbacher and S J. Derry (Eds.) Proceedings of the Twentieth Annual Conference of the Cognitive Science Society. Mahwah, NJ: Lawrence Erlbaum Associates

Elman, J. (1999). Generalization, rules, and neural networks: A simulation of Marcus et al., in press (http://crl.ucsd.edu/~elman/Papers/MVRV simulation.html)

Hadley, R F., Arnold, D., and Cardei, V. (1998). Syntactic Systematicity Arising from Semantic Predictions in a Hebbian-Competitive Network, Proceedings of the Twentieth Annual Conference of the Cognitive Science Society, Madison, Wisconsonsin: Lawrence Erlbaum Assoc., Publishers.

Marcus, G. F. (1998). Rethinking eliminative connectionism, Cognitive Psychology, vol. 37.

Marcus, G. F., Vijayan, S., Rao, S. B., Vishton, P. M. (1999). Rule learning in seven-month-old infants, Science, 283, 77-80.

Niklasson, L. F. and van Gelder, T. (1994). On being systematically connectionist. Mind and Language, 9:288-302

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. Cognition, 48, 21-69

A Computational Model of Counterfactual Thinking: The Temporal Order Effect

Clare R. Walsh (cwalsh@tcd.ie)
Psychology Department, University of Dublin, Trinity College,
Dublin 2, Ireland

Ruth M.J. Byrne (rmbyrne@tcd.ie)
Psychology Department, University of Dublin, Trinity College,
Dublin 2, Ireland

Abstract

People generate counterfactual alternatives to reality when they think about how things might have happened differently, 'if only…'. There are considerable regularities in the sorts of past events that people mentally undo, for example, they tend to mentally undo the most recent event in an independent sequence. Consider a game in which two contestants will win £1000 if they both pick cards from the same color suite. The first player picks black and the second red and they lose. Most people spontaneously undo the outcome by thinking, if only the second player had picked black. We describe a computational model that simulates our theory of the mental representations and cognitive processes underlying this temporal order effect. The computer model is corroborated by tests of the novel predictions of our theory: it should be possible to reverse the temporal order effect by manipulating the way in which the winning conditions are described.

Counterfactual Thinking

When people reflect on past events, they tend to think not only about the events that actually happened but also about how those events might have happened differently. This tendency to construct imaginary alternatives to reality is called counterfactual thinking (e.g., Kahneman & Tversky, 1982). For example, if your car breaks down and you are late, you might think that you would have been on time if you had had the car serviced or if you had taken the train.

Counterfactuals have been studied in philosophy (e.g., Lewis, 1973; Stalnaker, 1968), psychology (e.g., Kahneman & Miller, 1986) and artificial intelligence (e.g., Costello & McCarthy, 1999; Ginsberg, 1986). Counterfactual thinking has been implicated in many aspects of cognition and emotion. It may play a role in formulating counterexamples in reasoning (Johnson-Laird & Byrne, 1991) and in formulating sub-goals in problem-solving (Ginsberg, 1986). Counterfactuals may allow us to learn (e.g., Roese, 1994). The sorts of counterfactuals that are useful to people may also be useful to learning algorithms in artificial intelligence systems (Costello & McCarthy, 1999). Counterfactual thinking has also been linked to a range of emotions and social judgements, including blame and regret, both in the laboratory and in real-life settings.

Yet little is known about the mental representations and cognitive processes that underlie the generation of counterfactuals. Our goal in this paper is to describe a computational model that simulates a theory of the processes underlying counterfactual thinking and some experimental results that corroborate this theory.

Psychological studies of the sorts of counterfactuals that people generate indicate considerable regularities, despite the infinite number of ways that past events could have happened differently (e.g., Kahneman & Miller, 1986). People are more likely to undo exceptional than routine events (e.g., Kahneman & Tversky, 1982), actions than inactions (e.g., Byrne & McEleney, 2000), controllable than uncontrollable events (e.g., Girotto, Legrenzi & Rizzo, 1991; McCloy & Byrne, 2000) and the first event in a causal chain (e.g., Wells, Taylor & Turtle, 1987). In this paper we will focus on one important factor that influences the mutability of an event, that is, its temporal order in relation to other events, to which we now turn.

The Temporal Order Effect

Research has shown that when a series of events are independent of each other, people tend to mutate the most recent event (Byrne, Segura, Culhane, Tasso & Berrocal, 2000; Miller & Gunasegaram, 1990; Spellman, 1997). Consider a game in which two individuals are given a shuffled deck of cards, and each one picks a card from their own deck. If the two cards they pick are of the same color (i.e., both red or both black), each individual wins £1,000. Otherwise, neither individual wins anything. John goes first and picks a red card from his deck; Michael goes next and picks a black card from his deck. Thus the outcome is that neither individual wins anything (from Byrne et al, 2000). Participants tend to undo the second event, e.g., if only Michael had picked red too, when they are asked

to imagine that one of the card selections came out differently so that the players won, and this finding has been termed the temporal order effect. In addition, the second player, Michael, is usually expected to experience more guilt and to be blamed more by John. This effect has also been demonstrated in a number of practical situations, such as in judgements of fairness in an exam context (Miller and Gunasegaram, 1990) and in ranking team performance in a league (Sherman & McConnell, 1986

One possible explanation is that causality is assigned by the relative change in probability before and after an event (Spellman, 1997) although this account cannot explain the shift in focus which arises when an explicit alternative to the first event is provided (Byrne et al, 2000). An alternative explanation is that the first event in an independent sequence may be relatively immutable because it is presupposed (Miller & Gunasegaram, 1990), acting as a background against which later events are perceived (Sherman & McConnell, 1996), and playing an important contextualising role in constructing a mental representation of a factual situation (Byrne et al., 2000). Our aim is to explain why the first event is presupposed or perceived as immutable, and to do so, we will focus on the mental representation not only of the facts, but also of the counterfactual alternatives to the facts.

People may understand the card scenario by constructing a set of mental models (Johnson-Laird & Byrne, 1991), that is, mental representations that correspond to the structure of the world, and their models may represent certain aspects of the factual situation explicitly:

John red Michael black Lose

where 'John red' represents 'John picked a red card', 'Michael black' represents 'Michael picked a black card', and 'Lose' represents the outcome (Byrne et al., 2000). They may generate their counterfactual models by mutating aspects of the factual model. The fully explicit set of models is as follows:

Factual:	John red	Michael black	Lose
Counterfactual:	John red	Michael red	Win
	John black	Michael black	Win
	John black	Michael red	Lose

where separate possibilities are represented on separate lines of the diagram, and the models may be annotated to keep track of their epistemic status (Byrne & Tasso, 1999; Johnson-Laird & Byrne, 1991). The temporal order effect indicates that people construct just a subset of the possible counterfactual models:

Factual:	John red	Michael black	Lose
Counterfactual:	John red	Michael red	Win

. . .

where the three dots represents an implicit model which may be fleshed out to be more explicit if need be. Our aim is to explain why people construct this counterfactual model more than others.

TempCounterFacts: A Computational Model of the Temporal Order Effect

We will describe a computational model, called TempCounterFacts, which simulates the primary tenets of our theory of the mental representation and cognitive processes underlying the temporal order effect in counterfactual thinking (see Walsh & Byrne, 2001, for details). The program is written in LISP and it takes as input a set of facts about the card selection game (e.g., John picked red and Michael picked black) and a description of the winning conditions (e.g., if they both pick red or both pick black they win) and it generates a counterfactual alternative about how the outcome could have turned out differently (e.g., they would have won if Michael had picked red).

We suggest that the counterfactual context, that is, the representation of the conditions leading to a counterfactual outcome, such as the conditions under which a contestant would have won or lost a card game, can in themselves provide an explicit alternative to a factual event. This possibility has not been explored systematically before. Our suggestion is that people imagine a counterfactual scenario by changing their model of the facts to fit with their model of the winning conditions. They may select the first element of the factual model, e.g., John picks red, and find a match for it in the winning models. They may consider only this model as a possible counterfactual and conclude that if Michael had picked red they would have won. We suggest that the generation of a counterfactual alternative is driven not only by the 'bottom-up' facts of the actual situation but also by 'top-down' expectations derived from the counterfactual context. The program consists of three suites of functions.

1: Representing Facts and Counterfactual Context

The program begins by constructing a representation of the facts in a FactModel:

Factual: John Red Michael Black Outcome Lose

It also takes as input a set of winning conditions and it constructs the Counterfactmodels or set of models of the winning conditions:

John Red Michael Red Outcome Win
John Black Michael Black Outcome Win

These initial models represent the conditions under which the protagonists can win. Because of working memory constraints people may represent as little information as possible (Johnson-Laird & Byrne, 1991). For example, they may not explicitly represent the conditions under which the protagonists can lose (Byrne, 1997).

The program constructs models for different conditions which may be specific events, e.g., 'John picks a red card' or not, e.g., 'both players pick a red card'. It accepts four connectives: and, or (unspecified disjunction), ori (inclusive disjunction) and ore (exclusive disjunction).

2: Matching Facts and Counterfactual Context

The program attempts to match the FactModel to the set of CounterfactModels. It selects the first fact, e.g., John picked a red card, and it attempts to find a match for this event in the CounterfactModels. If an explicit match is found then it selects this model. If not, then it looks for a model which contains the negation of the first fact, e.g., John picks not-red. In the current example, it finds a match in the first model of the CounterfactModels:

John Red Michael Red Outcome Win

Once a match is found the program checks to see if the selected model is fully explicit, that is, it contains explicit information about both card selections. If the selected model is not fully explicit then the program fleshes out the models to be explicit (for details on fleshing out models and other technicalities, see Walsh & Byrne, 2001.)

The program then compares the FactModel and the selected CounterfactModel. If they match entirely then the FactModel is a winning instance. If there is only one item that is different then it uses this model to generate a new CounterAlternative model. If there is more than one difference, then the program continutes to search through the remaining CounterfactModels to find one which is more similar to the FactModel. In this way, the program ensures that minimal changes are made (see, e.g., Byrne, 1997).

3: Generating a Counterfactual Alternative

Once a counterfactual model has been selected, the program identifies the events in this model which are different from the FactModel. In the current example, it is Michael's card. It then generates a new CounterAlternative model by taking the FactModel and replacing the FactModel events with the CounterfactModel events, i.e., Michael picked Black is replaced with Michael picked Red. It describes the newly constructed CounterAlternative by generating a counterfactual conditional of the following form:

If it had been the case that: (Replaced event)
then it would have been the case that:
(Outcome Win).

The program simulates the temporal order effect, that is, it mutates the second event, when it is given the scenario in the current example, which is typical of scenarios used in such studies. However, the program also produces a novel reversal of the temporal order effect when it is given certain descriptions, as we will now describe.

Performance of the model on novel descriptions

We tested the performance of the model on several novel scenarios, with different sorts of winning conditions and different sorts of facts (see Walsh & Byrne, 2001, for full details). For example, we gave the model descriptions of a card game in which the winning conditions required the players to pick *different* colour cards and the fully explicit set of models for the winning conditions were as follows:

John Red Michael Black Win
John Black Michael Red Win

In each case, we presented the program with the same facts, i.e., John picks a black card and Michael picks a black card and they lose, and it produces the FactModel:

Factual: John black Michael Black Outcome Lose

We varied the way in which we described the winning conditions. Given the following 'black' disjunction, to describe the winning conditions:

If one or the other but not both pick a card from a black suit, each individual wins £1,000

the program constructs the following initial models:

John Black Win
Michael Black Win

and produces the temporal order effect. When one of the CounterfactModels contains an explicit match for the first fact, as in the models of the 'Black' disjunction the program selects this model. If it is not fully explicit then the program fleshes it out, relying on the footnotes to indicate how to do so. The program compares the fleshed out model to the FactModel and if the second event is different, as it is in the example, then the program uses this event to generate a new CounterAlternative model and to produce the counterfactual conditional:

If it had been the case that: (Michael not-Black)
then it would have been the case that:
(Outcome Win).

Given instead the following 'red' disjunction, to describe the winning conditions:

If one or the other but not both pick a card from a red suit, each individual wins £1,000

The program produces the following initial models:

John Red Win
Michael Red Win

and it produces a *reversal* of the temporal order effect, that is, it constructs a counterfactual scenario that undoes the first event rather than the second event. When the CounterfactModels do not contain an explicit match for the first fact, as in the models of the 'Red' disjunction, the program selects instead a model which contains the negation of the first fact, John picks not-black (which in the binary context of the color card

game, the program recognises as red). It repeats the same process described above, however in this case it is the first event which is different in the FactModel and CounterfactModel. As a result, this event is used to generate the CounterAlternative model and the conditional:

If it had been the case that: (John Red)
then it would have been the case that:
(Outcome Win).

The winning conditions are identical in both descriptions, and the facts are identical, but the description differs. As a result, the program constructs different sorts of initial models. The program simulates the assumption of the model theory that reasoners rarely construct a fully explicit set of models and their initial set of models makes some information explicit and some implicit (see Johnson-Laird & Byrne, 1991). One novel prediction of our theory is that it should be possible to reverse the temporal order effect if the way in which the winning conditions are described ensures that people construct an initial model that does not contain an explicit match to the first event. We turn now to some experimental results that corroborate this novel prediction.

Experimental Results on Temporal Order

We constructed a scenario based on the color card scenario (from Byrne et al., 2000). In a series of experiments, the facts of the players' selections remained the same: John goes first and selects a black card, Michael goes second and the card that he selects is also black, and the outcome is that both players lose. The winning conditions were also identical in each of the experiments: the players would win if they each picked different cards. We varied the description of these winning conditions (see Walsh & Byrne, 2001, for details). The experiments test our predictions that people hold this *counterfactual context* in mind from the outset and they use it to help them construct an appropriate counterfactual model.

In one experiment, we described the winning conditions as a disjunction of red cards:

If one or the other but not both pick a card from a red suit, each individual wins £1,000

and we compared it to a control description designed to eliminate the temporal order effect (see Walsh & Byrne, 2001 for details). In the experiment, the facts were that John picked black and Michael picked black and so they both lost. The two descriptions referred to exactly the same set of winning conditions, but for the 'red' disjunction, people cannot readily match the fact about the first player, John picked black, to their explicit thoughts about how the players can win. Instead the availability of an explicit alternative to the first fact, should lead them to mutate the first fact.

We tested 148 undergraduate students from different departments in the University of Dublin, Trinity College in several large groups. They took part voluntarily and were randomly assigned to the control or 'red' disjunction condition in a between subjects design. (Five participants were eliminated because they failed to follow the instructions.) Participants completed the following counterfactual mutation task:

> Please complete the following sentence. John and Michael could each have won £1,000 if only one of them had picked a different card, for instance if...

followed by several other related tasks (for details see Walsh & Byrne, 2001).

As our theory predicts, and as our cognitive model simulates, the 'red' disjunction reversed the temporal order effect. As Table 1 shows, the results indicated that for participants who mutated a single event, more mutated the first event (40%) than the second event (24%), and this difference was reliable (binomial n = 61, z = 1.79, 1-tailed $p < .04$), whereas when the same set of winning conditions were described in the control condition, the temporal order effect was eliminated. In the control condition, as many participants mutated the first event (32%) as the second event (36%, binomial n = 32, z = .18, p = .86), as we had expected (see Walsh & Byrne, 2001, for further details).

Table 1: The percentages of mutations in Experiment 1

	Control (n = 47)	Disjunction (n = 96)
Mutations		
First only	**32**	**40**
First and then Second	15	24
Second only	**36**	**24**
Second and then First	2	1
Neither	15	11

The experiment provides the first demonstration that the typical temporal order effect can be reversed, that is, participants mutate the first event in the sequence, rather than mutating the second event. The reversal depends not on the facts or on the nature of the winning conditions but on the way the winning conditions are described. Our explanation is that this description makes some information explicitly available in the mental models that reasoners construct, and renders other information implicit in the representation. An alternative to the first player's choice was made explicitly available and as a result, the temporal order effect was reversed.

Is it possible that the results show simply that the temporal order effect does not occur when the players

must pick different cards? The original temporal order effect may be an artifact of the constraint that both players must choose the same card. However in a further experiment in this series, we ruled out this possibility (see Walsh & Byrne, 2001). We showed that the temporal order effect can be observed even when players must pick different cards. We used the same scenario as described in the experiment here except that we changed the conditionals. We described the winning conditions as a disjunction of black cards:

If one or the other but not both pick a card from a black suit, each individual wins £1,000

Participants given this 'black' disjunction exhibited the standard temporal order effect. For both the 'red' and the 'black' disjunction conditions, the facts were the same: Both players picked black. The winning conditions were also the same (the contestants would have won if they picked different colored cards). The logical form of the description was the same, in that it was an exclusive disjunction. The only difference was in the reference to the color of the suit, black or red. This small difference in wording created a large difference in mutation patterns: mutations of the first event versus mutations of the second event. Our explanation is that reasoners represent the winning conditions not in a fully explicit set of models but in an initial set of models that makes some information explicit and keeps some implicit. We have called this mental representation of the winning conditions, the counterfactual context.

General Discussion

This paper provides one of the first computational simulations of counterfactual thinking. The model simulates our theory of the mental representations and cognitive processes that underlie counterfactual thinking, in the domain of the temporal order effect. One widely held view is that the mental representation of the facts are important in the generation of counterfactual alternatives. Our model makes use not only of the representation of the facts, but also of the representation of the winning conditions, which we have called the *counterfactual context*. It constructs representations that make some information explicit and leave other information implicit.

The program simulates the robust temporal order effect. However, our theory also led to a novel prediction about the reversal of the temporal order effect. In a series of experiments, we corroborated the predictions (see Walsh & Byrne, 2001). Our experiments showed that the temporal order effect can be reversed, eliminated or observed. The experiments provide the first demonstration that the temporal order effect can be reversed and that the nature of the description of the winning conditions can influence the mutability of events.

Acknowledgements

We thank Phil Johnson-Laird, Mark Keane, Orlando Espino, David Mandel, Rachel McCloy and Alice McEleney for their advice. The research was supported by Enterprise Ireland, the Irish Council for the Humanities and Social Sciences, and Dublin University. Some of the results were presented at the International Conference on Thinking in Durham, 2000.

References

Byrne, R. M. J. (1997). Cognitive processes in counterfactual thinking about what might have been. In D. L. Medin (Ed.). *The psychology of Learning and Motivation*, Vol 37. San Diego, CA: Academic Press.

Byrne, R. M. J. & McEleney, A. (2000). Counterfactual thinking about actions and failures to act. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Byrne, R. M. J., Segura, S., Culhane, R., Tasso, A., & Berrocal, P. (2000). The temporality effect in counterfactual thinking about what might have been. *Memory and Cognition, 28,* 264-281.

Byrne, R. M. J. & Tasso, A. (1999). Deductive reasoning with factual, possible and counterfactual conditionals. *Memory and Cognition,* 27, 726-740.

Costello, T., & McCarthy, J. (1999). Useful Counterfactuals. *Electronic Transactions on the Web.* Under Submission.

Ginsberg, M. L. (1986). Counterfactuals. *Artificial Intelligence*, 30, 35-79.

Girotto, V., Legrenzi, P., & Rizzo, A. (1991). Event controllability in counterfactual thinking. *Acta Psychologica*, 78, 111-133.

Johnson-Laird, P. N. & Byrne, R. M. J. (1991). *Deduction*. Hove, UK: Erlbau.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, 93, 136-153.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky, (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 201-208). New York: Cambridge University Press.

Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.

McCloy, R. & Byrne, R.M.J. (2000). Counterfactual thinking and the controllability effect. *Memory & Cognition.*

Miller, D. T. & Gunasegaram, S. (1990). Temporal order and the perceived mutability of events: Implications for blame assignment. *Journal of Personality and Social Psychology,* 59, 1111-1118.

Roese, N. J. (1994). The functional basis of counterfactual thinking. *Journal of Personality and Social Psychology*, 66, 805-818.

Sherman, S. J., & McConnell, A. R. (1996). Counterfactual Thinking in Reasoning. *Applied Cognitive Psychology*, 10, 113-124.

Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General,* 126, 323-348.

Stalnaker, R.C. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory*. Oxford: Basil Blackwell.

Walsh, C. and Byrne, R.M.J. (2001). A computational and experimental investigation of counterfactual thinking. *Manuscript in submission.*

Wells, G. L., Taylor, B. R. & Turtle, J. W. (1987). The undoing of scenarios. *Journal of Personality and Social Psychology*, 53, 421-430.

The Semantic Modulation of Deductive Premises

Clare R. Walsh (cwalsh@tcd.ie)
Psychology Department, University of Dublin, Trinity College,
Dublin 2, Ireland

P.N. Johnson-Laird (phil@princeton.edu)
Department of Psychology, Green Hall, Princeton University,
Princeton, NJ 08544, USA

Abstract

Two experiments examined how the mental models of premises influence deductive reasoning. Experiment 1 showed that individuals draw different conclusions from the same information depending on whether it is expressed in conditional assertions or disjunctions. It also showed that co-reference within the premises can speed up more difficult inferences. Experiment 2 corroborated these results and also showed that the failure to represent what is false can lead people to draw illusory inferences, i.e., systematic but compelling fallacies.

Introduction

The ability to reason deductively is central to human intelligence. Our goal is to examine how the nature of mental representations can influence this ability. Several factors should influence the process. First, the verbal formulation of premises should lead to different representations of the same underlying information. In turn, these representations should lead reasoners to draw different conclusions. Second, the semantic content of premises should influence the process of deductive reasoning. It should eliminate certain possibilities, and, in the case of co-reference from one clause to another, it may yield more concise representations. In this paper, we examine how these different factors impact on deductive reasoning.

Mental Models and Deductive Reasoning

The theory of mental models postulates that individuals who have no training in logic represent the meaning of assertions in mental models (Johnson-Laird and Byrne, 1991). Each mental model represents a possibility. But, the limitations of working memory force individuals to abide by the *principle of truth:* mental models represent only true possibilities and within them the constituent propositions in premises only when they are true. For example, an exclusive disjunction such as:

Either Mary is in Brussels or Gino is in Rome, but not both.

elicits two mental models of the alternative possibilities:

Mary-Brussels

Gino-Rome

where "Mary-Brussels" denotes a model of Mary in Brussels, and "Gino-Rome" denotes a model of Gino in Rome. Mental models of possibilities do not represent the falsity of the clauses in the disjunction, e.g., that Gino is not in Rome in the first possibility. One consequence is that mental models should give rise to illusory inferences, i.e., fallacies that most individuals make. Recent studies have corroborated their occurrence (e.g. Johnson-Laird and Savary, 1999). In one experiment, participants were given the following problem:

If there is a king in the hand then there is an ace in the hand, or else if there is not a king in the hand then there is an ace in the hand.
There is an king in the hand.
What follows?

All the participants drew the conclusion that there is an ace in the hand. Mental models yield this conclusion, but it is wrong. In fact, the sentential connective or else implies that one of the conditionals could be false. But, if the first conditional is false, then there isn't an ace in the hand even though there is a king in the hand.

The meanings of clauses, their co-referential relations, and background knowledge, can all *modulate* the basic meanings of sentential connectives (Johnson-Laird and Byrne, 2001). One way modulation can occur is when background knowledge prevents the construction of a model. Consider, for example, the conditional:

If she played a game, then she didn't play soccer.

The basic interpretation of a conditional allows the possibility that she doesn't play a game but that she plays soccer. This possibility, however, is eliminated by the knowledge that soccer *is* a game.

A second example of modulation, in our view, is Bouquet and Warglien's (1999) discovery that

reasoning from disjunctions was more accurate when the clauses were co-referential, e.g.:

Either Gino is in Brussels or Gino is in Rome, but not both. Gino is not in Rome. What follows?

Premises of this sort yielded a greater proportion of valid conclusions, e.g.: Gino is in Brussels, than premises that did not refer to the same individual in both clauses. Co-reference may enable people to construct representations that are more concise. The aim of Experiment 1 was to examine this possibility.

Experiment 1: Co-reference in reasoning

Our conjecture is that co-reference may allow reasoners to build a more concise representation of the premises, thereby reducing the load on working memory. We predicted that such representations should facilitate performance in reasoning tasks particularly if the demands on working memory are high. In addition, the form of the premises - whether they are based on a disjunction or a conditional - should make different information available. Hence, each sort of premise should make some inferences easier and some more difficult to draw.

Each participant carried out 32 inferences in a different random order. The inferences concerned people carrying out various actions. For half the inferences, the first premise was an exclusive disjunction, e.g.:

Rachel is climbing up the stairs or David is cooking at the stove but not both

which should elicit the mental models:

Rachel-climbing

David-cooking

For half the inferences, the same information was expressed as a biconditional, e.g.:

If and only if Rachel is not climbing up the stairs then David is cooking at the stove.

which should elicit the mental models:

¬ Rachel-climbing David-cooking

. . .

where "¬" denotes negation, and the ellipsis denotes an implicit model, which acts as a "place holder" for the possibilities in which the antecedent is false. If necessary, it can be fleshed out explicitly.

The second premise was a categorical assertion or categorical denial of either the first or second proposition in the preceding premise. There were accordingly 8 forms of inference.

In order to manipulate co-reference, there were four types of semantic content:

1. Two persons do two different actions (two-actions).
2. Two persons do one action, which they cannot perform simultaneously (exclusive-action, e.g. "sit on the stool").
3. Two persons do one action, which they can perform simultaneously (inclusive-action, e.g., "sit on the sofa").
4. One person does two different actions, which cannot be performed simultaneously (one-person).

The eight forms of inferences with the four sorts of content yielded the 32 different inferences.

The problems were presented on a computer. The participants drew their own conclusion about what followed from the premises. They responded by typing their answer, and their latencies were measured from the presentation of the premises to the first key press. They were not told that their responses were being timed. We tested 30 undergraduates from Princeton University in return for course credit.

Results and Discussion

Table 1 presents the percentages of correct responses to the eight forms of inference (collapsing over their contents). As the Table shows, the participants were more accurate when the categorical premise matched an event that was represented explicitly in the models of the first premise. When the categorical premise was negative and concerned the first event in the preceding premise, Not-A, the participants were more accurate in reasoning from the biconditional (91%) than the disjunction (68%). But, when the categorical premise was affirmative, A, they were more accurate in reasoning from the disjunction (94%) than from the biconditional (78%; McNemar tests, chi-squared = 17.36 $p < .0001$; chi-squared = 15.04, $p < .0001$ respectively). Participants performed equally well on both descriptions when the categorical premise affirmed the second proposition, B, (94% for both, McNemar Test, chi-squared = 0.07 $p = 1$). We predicted these results in terms of mental models, but they might reflect the surface matching of clauses in the premises. One result, however, is more readily explained in terms of models. When the categorical premise negated the second proposition, not-B, participants were more accurate in reasoning from the disjunction (68%) than from the biconditional (46%; McNemar Test, chi-squared = 13.8, $p < .0002$). Not-B mismatches the clauses in both sorts of premises. It is also not represented explicitly in the initial models of either premise. But, reasoners may find it easier to flesh out the disjunction, which already contains two mental models, than to flesh out the conditional which is represented by just one explicit mental model.

Table 1
The percentages of correct responses to the eight forms of inference in Experiment 1

	First Premise	
Categorical Premise	*Iff not-A then B*	*A or else B*
A	78	94
Not-A	91	68
B	94	94
Not-B	46	68

The co-referential manipulation had no effect on accuracy. Yet, it did affect the latencies of correct responses. The principal results were that responses to exclusive-action problems (10.93 secs) and to inclusive-action (10.18 secs) were faster than those to two-action problems (11.58 secs; Wilcoxon test $z = 2.33, 2.81$, $p < .01$, $p < .003$, respectively, excluding the results of two outliers). In other words, co-reference can help the process of reasoning. In particular, problems in which two persons carry out one action, whether or not they can perform it at the same time, elicited faster responses than problems in which two persons carry out two different actions. Reference to a common action may yield more parsimonious models of the premises, and it may help reasoners to avoid confusion about which action a particular individual was carrying out. Such confusions are more likely in the case of a disjunction, which, unlike a biconditional, demands that reasoners model two explicit possibilities. The shared referent in a disjunction is common to two alternative models whereas in a biconditional it occurs within one model. That, perhaps, is why the referential effects were stronger for disjunctions (exclusive-action 9.0 secs vs two-action problems 10.7 secs, inclusive-action 8.7 secs vs two-action problems 10.6 secs, Wilcoxon test $z = 2.76, 2.44$, $p < .003, .01$ respectively), than biconditionals (exclusive-action 11.8 secs vs two-action problems 12.3 secs, inclusive-action 11.7 secs vs two-action problems 12.0 secs, Wilcoxon test $z = 0.68, 1.47$, $p < .25, .07$ respectively). We followed up these phenomena in a second experiment.

Experiment 2: Illusory inferences and co-reference

The aims of the experiment were twofold. The first aim was to examine what inferences people make from an exclusive disjunction of the form:

Either P and Q or otherwise R and S.

The disjunction was paired with a categorical premise, either asserting or denying its first proposition, P. The participants had to evaluate the validity of a one-clause conclusion, either Q or R. There were accordingly four forms of inference.

The model theory predicts that the failure to make certain information explicit in the models of the disjunction should lead people to make invalid inferences. According to the principle of truth, reasoners should construct two mental models of such a premise:

P Q
R S

It follows that given the categorical premise P, reasoners should infer that Q and not-R follow. Similarly, given the categorical premise not-P they should infer that not-Q follows. These inferences, however, are illusions. When falsity is taken into account, the disjunctive premise is consistent with six different possibilities, which we present here in *fully explicit* models:

P Q ¬ R S
P Q R ¬ S
P Q ¬ R ¬ S
¬ P Q R S
P ¬ Q R S
¬ P ¬ Q R S

These models show that the three previous inferences are invalid. Given the premise, P, for instance, Q and not-Q are both possible, and likewise R and not-R are both possible. In contrast, given the categorical premise, not-P, participants should correctly infer that R follows: this conclusion follows from the mental models above, but it also follows from the fully explicit models.

Our second aim was to examine the effects of co-reference on these inferences. Experiment 1 showed that co-reference reduced response times, at least for certain inferences. The present experiment followed up this effect and the semantic modulation of the premises.

Table 2 presents the five sorts of semantic contents, which manipulated the number of shared referents (i.e., people carrying out actions) contained in the first premise, and whether the co-referential relations occurred within or between models. We predicted that co-reference would again facilitate performance and that this facilitation would be greater as the number of shared referents increased. We also predicted that facilitation would be greater for problems requiring a greater working memory load.

Table 2: The five sorts of semantic content in Experiment 2, and the number of fully explicit models compatible with each content.

1. *Four persons act: six models*

Either Jane is kneeling by the fire and Sean is looking at the TV or otherwise Mark is standing at the window and Pat is peering into the garden.

2. Two persons, one per clause: six models

Either Jane is kneeling by the fire and she is looking at the TV or otherwise Mark is standing at the window and he is peering into the garden.

3. Two persons do inclusive actions: six models

Either Jane is kneeling by the fire and Mark is standing at the window or otherwise Jane is looking at the TV and Mark is peering into the garden.

4. Two persons do exclusive actions: two models

Either Jane is kneeling by the fire and Mark is looking at the TV or otherwise Jane is standing at the window and Mark is peering into the garden.

5. One person: two models

Either Jane is kneeling by the fire and she is looking at the TV or otherwise she is standing at the window and she is peering into the garden.

The first three sorts of content in Table 2 are consistent with all six of the fully explicit models above. However, in the other two cases, the exclusive actions rule out the models in which P and R occur together, and Q and S occur together. The premise is therefore consistent with just two fully explicit models:

P	Q	¬ R	¬ S
¬ P	¬ Q	R	S

These models yield the same conclusions as the mental models described above, but these conclusions are no longer illusions, but correct.

We tested individually 35 participants (25 paid members of the public and 10 postgraduate volunteers from the University of Dublin, Trinity College). They acted as their own controls and carried out the 20 inferences in different random orders. We constructed 20 sets of materials, each of which contained the same number of words. The materials were rotated so that they were presented equally often with each of the five sorts of semantic content.

The problems were presented on a computer screen. For each problem, the premises and questions were presented one-by-one on the screen. Participants responded to the question by pressing one of three keys: yes, no or cannot tell. The program recorded separately the time that it took participants to read each of the premises and to answer the question. The participants were not told that their responses were being timed.

Results and Discussion

Table 3 presents the percentages of the "Yes" and "No" responses to the main sorts of problems. As the table shows, the participants succumbed to the illusory (six model) problems (10% correct), but performed well the (six-model) control problem (78% correct). Of the 35 participants, 34 were less accurate on the illusions than on the control problems (Sign Test, $p < 1$ in 900 million). The pattern of responses was comparable for the two model problems, and so we suspect that the participants constructed just two models for the six-model and the two-model problems. A similar pattern of results occurred in the participants' responses to the first problem that they encountered. They were more accurate with the two-model problems (81% correct) than with the six-model problems (31% correct). This result suggests that the illusions occurred spontaneously and not as a result of the development of a mental set.

Table 3: The percentages of the "Yes" and "No" responses in Experiment 2. The balance of the responses (around 10% per problem) were "can't tell". The predicted responses are shown in bold, and underlined where they are illusory.

	Six models		Two models	
1. P & Q, or R & S	Yes	No	Yes	No
P				
∴ Q	**87**	4	**84**	6
2. P & Q, or R & S				
P				
∴ R	22	**69**	7	**87**
3. P & Q, or R & S				
Not-P				
∴ Q	13	**75**	12	**77**
4. P & Q, or R & S				
Not-P				
∴ R	**78**	15	**83**	10

There was no reliable difference between the five semantic conditions in response accuracy or in the time that it took to read the first premise. But, a significant difference in response times occurred across conditions when all responses were considered (Friedman Test, chi-squared = 20.08, df = 1, $p = .0005$), and when we included only the responses predicted by the mental model theory (Friedman Test, chi-squared = 10.66, df = 1, $p = .03$). All further analyses are based only on the responses predicted by the mental model theory, because it is difficult to know what the participants were doing when they got the answer right to the illusory problems and wrong to the control problems.

There were two principal results:

1. Responses were faster when the first premise referred to fewer individuals (1-person condition, mean 6.8 secs, 2-person conditions, mean = 8.6 secs, and 4-person condition, mean = 8.3 secs; Kruskall Wallis Test, chi-squared = 10.42, df = 2, $p < .003$).

2. The difference was significant in the different-model condition (1-person condition, mean 7.4 secs, 2-person conditions, mean = 9.9 secs, and 4-person condition, mean = 10.5 secs; Kruskall Wallis Test, chi-squared = 12.2, df = 2, $p < .002$) but not in the same-model condition (1-person condition, mean 6.3 secs, 2-person conditions, mean = 7.4 secs, and 4-person condition, mean = 6.6 secs; Kruskall Wallis Test, chi-squared = 1.89, df = 2, $p < .20$).

The results corroborated the occurrence of illusory inferences, and reasoners seem likely to construct just two models of disjunctions of the form:

Either P and Q or otherwise R and S.

They overlook the different ways in which the conjuncts could be false in the case of the six model problems, i.e., those with a content that does not eliminate any of the possibilities. The latencies of the responses bear out our conjecture that inferences are easier when the premises concern fewer individuals. Reasoners can construct more concise models in this case and are less open to confusion about who did what. Faster responses occurred both when the co-referential relation was within one model and when it was between items in different models. However, response times were faster only when reasoners had to consider an alternative model to the one referred to in the categorical premise, probably because that condition places an extra load on working memory.

General Discussion

The mental model theory predicts that when people represent a premise, they do so in accordance with the principle of truth. They construct a representation that makes only some information explicit. Experiment 1 corroborated the principle. It showed that reasoners draw different conclusions from given information, depending on whether it is expressed as a biconditional or an exclusive disjunction. Likewise, Experiment 2 showed that the failure to represent falsity can lead reasoners to make illusory inferences.

A second factor influences mental models: the occurrence of co-reference within the premises. Experiment 1 showed that co-reference within an exclusive disjunction speeded up the process of inference. Experiment 2 showed the same effect, but only when reasoners had to consider an alternative model to the one referred to in the categorical premise. The inferential task in such cases places a bigger load on working memory, and co-reference evidently ameliorates matters. The same factor may explain why there was no facilitation for conditionals in our first experiment. Reasoners can construct a more concise representation of premises containing co-referents. This parsimony reduces the load on working memory and the latencies of more difficult inferences.

Acknowledgements

Clare Walsh is supported by an Enterprise Ireland Ph.D. Fellowship and a Government of Ireland Scholarship for the Humanities and Social Sciences, and a Dublin University Postgraduate Award. We thank Ruth Byrne, Eugenia Goldvarg, Uri Hasson, Markus Knauff, Yingrui Yang, and Lauren Ziskind, for their helpful comments on this research.

References

Bouquet, P. & Warglien, M. (1999). Mental Models and Local Models Semantics: the Problem of Information Integration. *Proceedings of the European Conference on Cognitive Science* (ECCS'99), pp.169--178, Siena, Italy.

Johnson-Laird, P.N. & Byrne, R.M.J. (1991). *Deduction*. Hove and Hillsdale: Erlbaum.

Johnson-Laird, P.N. & Byrne, R.M.J. (2001). Conditionals: a theory of meaning, pragmatics and inference. *Submitted.*

Johnson-Laird, P.N, & Savary, F. (1999). Illusory inferences: A novel class of erroneous deductions. Cognition, 71, 191-229.

The Appearance of Unity: A Higher-Order Interpretation of the Unity of Consciousness

Josh Weisberg (jwsleep@aol.com)
CUNY Graduate Center
Department of Philosophy, 365 5th Avenue
New York, NY 10016 USA

Abstract

Recent developments in neuroscience and psychology have put pressure on the traditional philosophical idea of the unity of consciousness. Some have argued that split-brain cases and multiple personality disorder demand a rejection and elimination of the very notion of a unified consciousness. In this poster, I argue that David Rosenthal's higher-order-thought theory of consciousness allows for an explication of unity that provides for the subjective appearance of unity, but respects the actual and potential disunity of the brain processes that underwrite consciousness.

Introduction

Researchers often have several distinct phenomena in mind when they discuss the unity of consciousness. One issue, which occupies a large part of the neuroscientific focus on unity, is the "binding problem." How are features processed in disparate anatomical and functional locations in the brain brought together into one coherent experience? Closely related is the problem of accounting for the apparent "spatial" relations that hold between our sensations, relations that allow us to view one feature as next to another in the same experience. A third phenomenon associated with unity is the apparent clear and seamless nature of our conscious experience. Finally, there are issues of ownership and the self. Why is it that my conscious perceptions belong, as it were, to me? What is the nature of this "I" that the perceptions belong to?

In this poster, I propose to address the issue of unity by first introducing a theoretical model of the conscious mind, David Rosenthal's higher-order-thought hypothesis, and then seeing which aspects of unity can be explained in terms of nonconscious neuroscientific or psychological processes. Then I will attempt to show how the remaining elements of unity can be adequately dealt with by the theory. I will close by briefly considering some worries about eliminativism that often accompany discussions of unity and consciousness.[1]

The HOT Theory

To begin, I will outline Rosenthal's higher-order-thought hypothesis. The view distinguishes between a transitive and an intransitive use of the term "conscious." The transitive use occurs when we talk about being conscious of something. The intransitive use applies to mental states, which can be described as conscious or nonconscious. The theory holds that for a mental state to be intransitively conscious, we must be *transitively conscious of it* in a suitable manner. Arguably, this process is best explained by the presence of a "higher-order thought" (HOT) to the effect that we are in that mental state. HOTs are intentional states that represent the subject as being in mental states. Being the target of a HOT is what makes a state conscious.

Furthermore, the HOT must arise in the appropriate way. To account for the seeming immediacy of conscious experience, we must be unaware of any inference or observation that causes the tokening of the HOT. Mental states not targeted by HOTs are not conscious states. so the HOTs themselves are generally not conscious states, unless targeted by an additional HOT. To summarize, the theory holds that mental states are conscious when we represent ourselves as being in those states, and this representation is achieved by higher-order thought.[2] .

---

[1] It is important to note that my work on this topic is strongly influenced by David Rosenthal's writings and instruction. Rosenthal has recently completed several articles concerning Unity (Rosenthal, forthcoming a, forthcoming b), and my work herein is meant to compliment that treatment.

[2] Obviously, much more can be said concerning the theory. I hope that this brief outline will suffice for the purposes at hand, and I will fill in details as they become relevant. For a full description and defense of the HOT theory, see Rosenthal

Lower-level Unity

Neuroanatomy tells us that the brain processes that underwrite perception are widely distributed. This raises the question of how the disparate elements of perception are combined into a single coherent experience. How does the brain bring it all together? This is the famous binding problem. The layered explanatory model of the HOT theory allows for this issue to be dealt with largely at the level of nonconscious processes. Promising work has been done in laying out a *temporal* solution to the binding problem (see, e.g., Crick and Koch, 1990; Llinás and Ribary, 1993; Edelman and Tononi, 2000). Very roughly, it is hypothesized that when groups of neurons oscillate at the appropriate rate, they bind by firing in synchrony. Arguably, this solution does not involve consciousness mental states at all.

Unconscious perceptions, like those registered in priming or subliminal perception, influence behavior in virtue of their perceptible properties, and these properties need to be bound, just as in conscious perception. In order for a priming image to have an appropriate semantic effect (for example, in influencing the disambiguation of a sentence), we must perceive the image's features as unified, in order to make the proper identifications. Perceptions are bound whether they are conscious or not. The processes that bind the percepts occur independently from consciousness, so this aspect of unity is not one that a theory of consciousness has to explicate.

Another aspect of unity, our awareness of experienced objects as located in this or that portion of a sensory field, can also be dealt with at the level of nonconscious processing. In vision, for example, various functional locations in the visual cortex nonconsciously extract and organize information concerning spatial orientation. However, this feature of unity can be more fully explicated in terms of sensory qualities. In any event, the explanation can proceed independently of consciousness.[3]

Sensory qualities are properties of mental states grouped in families that exhibit a structural relationship with families of perceptual properties in the world. Our commonsense taxonomy of perceptible color properties, for example, includes similarity and difference relations. Red is more similar to orange then it is to green. The sensory states employed in color vision must maintain these relationships. They must be in families with similarities and differences homomorphic to the similarities and differences that hold among perceptible properties. The spatial properties of sensory states can be dealt with in the same way. Sensory states that underwrite visual spatial perception must possess similarities and differences homomorphic to those possessed by perceptible spatial properties.[4]

In this fashion, sensory qualities can be explained independently of consciousness. What it is for a sensation to be located next to, or above, or below, another sensation in conscious awareness can be explicated by referring to the sensory qualities already possessed by those states. States have the features that locate them in mental space independently of their being conscious. So this feature of unity can be explained without invoking a theory of consciousness.

Thus, these elements of unity, binding and location in a sensory field, can be dealt with without reference to conscious mental states. Our perceptions have these features independently of consciousness. When we do become conscious of perceptual states, our HOTs represent them in terms of these antecedently present features. The act of becoming conscious of these features does not bring them into being. Our mental states are bound and organized in sensory fields independently of consciousness. Our HOTs simply represent these states as they are: bound, unified percepts, occupying this or that portion of the relevant sensory field. In this way, we become conscious of a bound, unified experience.

Conscious Unity

However, even though nonconscious perceptual processes bind and organize our perceptual sensations, it is clear that in some respects our conscious awareness of the world outstrips what is delivered by our perceptual mechanisms. This is exemplified by Daniel Dennett's discussion of a conscious perception of wallpaper made up of many repeated images of Marilyn Monroe (Dennett, 1991, pg. 354-355). When standing in front of such an array, we seem instantly to see that the wallpaper is all Marilyns, and what's more, we see this with an apparent clarity and depth that seems to take in the whole scene with equal acuity. But appearances may be deceiving here. Outside of the small area of foveal vision, we do not actually process visual stimuli with the fine-grainedness that we are aware of in consciousness. But in conscious experience, we seem to be aware of a full, rich, clear, unified expanse. How are we to account for this?

Here, we can refer to the clarifying effect of HOT to explain the appearance of unity. According to the HOT theory, what our HOTs represent us as

---

1986, 1991, 1993, 1997, 1999b. For critical views, see Byrne, 1997; Dretske, 1995; Stubenberg, 1998.

[3] Rosenthal, 1991.

[4] See Rosenthal, 1991, 1999a, 1999b; Sellars, 1956; Shoemaker, 1975; Clark, 1993.

experiencing is what we consciously experience. Our nonconscious perceptual processes deliver a range of information about the visual scene, which grades off sharply as we move away from foveal vision. But in the absence of any alarming discontinuity detectable outside of the foveal area, our HOTs can simply represent the scene in a "more of the same" manner. HOT can represent to the effect that we are in a perceptual state with such-and-such sensory qualities, and those features repeat in a clear and unbroken pattern to the edge of the visual field.

It might be argued that if this is the case, then the homogenizing affect of the HOTs should be readily noticeable in consciousness. However, we will not ordinarily notice the smoothing over imposed by HOT because whenever we look more closely at a scene to see if it really is unbroken and repeating, we refocus our attention, and token a more detailed, though less broad, HOT. Then we become conscious of that particular detail, but are no longer conscious of the whole scene. So we won't be aware of the cleaning-up effect of the HOT when we try to attend to it. But given the *apparent* clarity and depth of the original perception of the Marilyns, and the absence of clear perception away from foveal vision, the smoothing-over effect of the HOT is the best explanation of our unified conscious experience.

Ownership and the Self

Finally, I will turn to the sorts of issues that most often arise in philosophical discussions of unity, the apparent fact that our perceptions are in a sense "owned" by us, and that this ownership points to the presence of a "self" to do the owning. Interesting evidence from neuroscience and psychology over the last several decades (split-brain cases, hemineglect, and multiple-personality disorder [MPD], for example) have put pressure on the idea that there really is a unified self in the brain which experiences all of our perceptions and thoughts. But it sure seems to us that we are unified individuals as we undergo conscious experience. How can we account for this, on the HOT model?[5]

First, we need a little more detail concerning the content of a HOT. A HOT represents to the effect that "I, myself, am in a state with such-and-such properties." By representing in this manner, employing the concept "I," the HOT ascribes the mental state to "I," the self. In doing so, the HOT serves to tag the conscious state to this self, and so the state is represented as being mine. In this way, I "own" the state.

[5] What follows is closely related to Rosenthal's recent treatment in (Rosenthal, forthcoming a, forthcoming b).

It is important to note the indexical features of the concept "I." "I" serves a function much like that of the term "here," which automatically refers to the location of the utterance, and gets its more specific content on the particular occasion its use from the location that it occurs in. "I" works in a similar way. When "I" occurs in a HOT, it refers back to the thinker that tokens the thought. It is the function of a HOT to pick out various states that the organism is in, and by employing the indexical "I" there is a thin sort of immunity to error present in the self-ascription by the HOT, reminiscent of the way that uses of "here" are immune to error by automatically referring to the location of utterance. Our conscious states are owned by us because in becoming conscious of them, we represent the states in conjunction with "I," which automatically refers back to the producer of the thought (cf. Shoemaker, 1968).

So, when we become conscious of a mental state, we token a HOT to the effect that we are in that state. This provides us with a sense of self. We are, so to speak, a creature that ascribes states as being present in that very creature, and ascribes them by employing the concept "I." In this way we become conscious of *our* states, belonging to us.

But more must be said about this "thinker" that "I" refers back to. The concept "I" is best seen as a theoretical term in the folk-psychological theory that enables us to ascribe mental states to ourselves and others. Folk psychology posits the self as the referent of "I," and it posits a variety of features of the self. The self takes in perceptions and initiates action. It has direct, infallible access to thoughts and sensations. It accounts for the continuity we have as individuals, and makes us who we are. But this notion of self has come under strong pressure from philosophy, psychology, and neuroscience. Is there anything left that can serve as the referent of the concept "I"?

I believe that several components of the folk-psychological notion of self can be preserved. First, there are the "biological" or "ecological" elements of the self. These are the features that allow us to navigate through the world, and to distinguish the boundaries of our bodies from the external environment. They are relatively low-level features of an organism's psychology. Primitive creatures like lobsters can respond differentially to their bodies and the environment, and so avoid eating their own legs. The processes that underwrite this ability can serve as the reference of "I." (See Dennett, 1996; Bermúdez, 1998.)

In more developed creatures like us, "I" also picks out those features that account for our psychological continuity. "I" refers to a collection of moods, memories, and abilities present in the individual. We possess a core group of states that define us as a subject. We have memories about who we are, various labels,

like a name, an address, a social security number, we have abilities, like the ability to play the guitar, or to drive, and we have deep seated moods and personality traits, like being lazy, mellow, or high strung.

This group of states shifts over time, and the boundaries between core states and more peripheral states is not a firm one. Furthermore, we may not be able to bring these states to consciousness with any clarity. We may in fact confabulate the content of the states as we become conscious of them, possibly altering their content as we do so. The self is like a novel that is constantly edited and revised as it is read. So the "I" present in HOT refers to those elements of an organism that allow it to negotiate through the environment and distinguish its body from the world, and it refers to the core psychological states that define us as an individual. (See Damasio, 1994, Chp. 10; Ramachandran and Blakeslee, 1998, Chp. 12.)

But what of the remaining elements of the folk-psychological notion of self? Why does it seem to us that we are free agents with direct, infallible access to the content of our own thought? Here, once again, the HOT that makes us aware of our states misleads. The concept of self that we employ in HOT is the unrevised folk-psychological one. It goes beyond what is really there in our minds. It seems to us that we are this sort of being because our folk-psychological concept posits such features, and our HOTs employ this concept in making us conscious of our thoughts and experiences. But on this score, we are in error.

So, some aspects of the folk-psychological conception of self which informs our intuitions about unity are mistaken. The self is not what it seemed to be. Does this entail that we don't really have selves, and that the unity of consciousness is a fiction, an illusion? I would argue that this worry rests on an overly-strong criteria of what marks a concept for elimination. In this case, we do maintain some aspects of our folk-psychological conception of self, namely its connection with autobiographical continuity and ability. We also can see how other aspects unity, like the presence of bound, spatially located perceptions "owned" by a subject, are maintained by our theory. I suggest that the self is still there, and that consciousness is indeed unified, but things are not exactly as they appeared prior to our investigations.

Acknowledgments

A version of this poster was presented as a paper at the ASSC 4 conference in Brussels, July, 2000. My thanks to the audience there for helpful comments. Thanks also to Roblin Meeks, Doug Meehan, and especially David Rosenthal.

References

Bermúdez, J. L. (1998). *The Paradox of Self-Consciousness*. Cambridge, MA: MIT Press.

Byrne, A. (1997). "Some Like it HOT: Consciousness and Higher-Order Thoughts." *Philosophical Studies* 24, 4 (Sept.) 1-27.

Clark, A. (1993). *Sensory Qualities*. Oxford: Clarendon Press.

Crick, F. H. C., and Koch, C. (1990). "Towards a Neurobiological Theory of Consciousness." From *Seminars in the Neurosciences*, 2, 263-275.

Damasio, A. (1994). *Descartes' Error*. New York: Avon Books.

Dennett, D. C. (1991). *Consciousness Explained*. Boston: Little Brown.

Dennett, D. C. (1996). *Kinds of Minds*. New York: Basic Books.

Dretske, F. (1995). *Naturalizing the Mind*. Cambridge: MIT Press.

Edelman, G. and Tononi, G. (2000). *A Universe of Consciousness*. New York: Basic Books.

Llinás, R. and Ribary, U. (1993). "Coherent 40-Hz Oscillation Characterizes Dream State in Humans." *Proceedings of the National Academy of Science*, 90, 2078-2081.

Ramachandran, V. S. and Blakeslee, S. (1998). *Phantoms in the Brain*. New York: Quill/William Morrow.

Rosenthal, D. M. (1986). "Two Concepts of Consciousness." *Philosophical Studies* 49, 329-359.

Rosenthal, D. M. (1991). "The Independence of Consciousness and Sensory Qualities." In Villanueva, ed. *Consciousness: Philosophical Issues, 1*. Atascadero, CA: Ridgeview Publishing Company.

Rosenthal, D. M. (1993). "Thinking that one Thinks." In M. Davies and G. W. Humphreys (Eds.) *Consciousness*. Oxford: Blackwell.

Rosenthal, D. M. (1997). "A Theory of Consciousness." in Block, N., Flanagan, O., and Guzeldere, G.,(eds.), *The Nature of Consciousness: Philosophical Debates*, Cambridge, MA: MIT Press, 729-753.

Rosenthal, D. M. (1999a). "The Colors and Shapes of Visual Experiences." In *Consciousness and Intentionality: Models and Modalities of Attribution*, D. Fisette (ed.), Dordrecht: Kluwer Academic Press, 95-118.

Rosenthal, David M. (1999b), "Sensory Quality and the Relocation Story," *Philosophical Topics*, 26, 1 and 2, (Fall and Spring), pp. 321-350.

Rosenthal, David M. (forthcoming, a), "Introspection and Self-Interpretation," *Philosophical Topics* (Winter 2001).

Rosenthal, David, M. (forthcoming, b), "Persons, Minds, and Consciousness," in *The Philosophy of Marjorie Grene*, in *The Library of Living Philsophers*, Hahn, L. E. (ed.), La Salle, Illinois: Open Court.

Sellars, W. (1956). "Empiricism and the Philosophy of Mind." *Minnesota Studies in the Philosophy of Science*, vol. 1, ed. Herbert Feigl and Michael Scriven, University of Minnesota Press, Minneapolis.

Shoemaker, S. (1968). "Self-Reference and Self-Awareness." *Journal of Philosophy*, 65/19, 555-567.

Shoemaker, S. (1975). "Functionalism and Qualia." *Philosophical Studies* 27: 291-315.

Shoemaker, S. (1991). "Qualia and Consciousness." *Mind* 100:4, 507-524.

Stubenberg, L. (1998). *Consciousness and Qualia*. John Benjamins Publishing Co., Amsterdam/Philadelphia.

How to Solve the Problem of Compositionality by Oscillatory Networks

Markus Werning (markus.werning@uni-erfurt.de)
Department of Philosophy, P.O. Box 900221
D-99105 Erfurt, Germany

Abstract

Cognitive systems are regarded to be compositional: The semantic values of complex representations are determined by, and dependent on, the semantic values of primitive representations. Both classical and connectionist networks fail to model compositionality in a plausible way. The paper introduces oscillatory networks as a third alternative. It provides neurobiological evidence for the adequacy of those networks and argues that they are compositional. Oscillatory networks combine the virtues and avoid the shortcomings of classical and connectionist architectures.

Compositionality and Systematicity

Minds have the capacity to compose contents. Otherwise, they would not show a systematic correlation between representational capacities: If a mind is capable of certain intentional states in a certain intentional mode (perception, thought, imagination, preference, etc.), it most probably is also capable of other intentional states with related contents in the same mode. The capacity to see a red square in a green circle, e.g., is statistically highly correlated with the capacity to see a red circle in a green square. To explain this empirical phenomenon, which is closely related to the well-known binding problem, compositional operations are postulated. They enable the system to build complex representations from primitive ones so that the semantic value of the complex representation is determined by, and dependent on, the semantic values of the primitives. Several theories have been developed to meet the requirement of compositionality. Both classical and connectionist attempts suffer from severe deficits, though.

Fodor and Pylyshyn (1988) for one take recourse to a language of thought, which they link to the claim that the brain can be modeled by a Turing-style computer. A subject's having a mental representation, they believe, consists in the subject's bearing a computational relation to a mental sentence; it is a relation analogous to the relation a Turing machine's control head bears to the tape. Accordingly, the mind composes complex representations from primitive ones just the way a computer composes phrases from words: by concatenation. The trouble with classical computer models is well known and reaches from the frame problem, the problem of graceful degradation, and the problem of learning from examples (Horgan & Tienson, 1996) to problems that arise from the content sensitivity of logical reasoning (Gigerenzer & Hug, 1992).

To avoid the pitfalls of classicism, connectionist models have been developed. Some of them attempt to meet the compositionality constraint. Smolensky (1995) maps the terms and the syntax of a language homomorphically onto an algebra of vectors and tensor operations. Each primitive term of the language is assigned to a vector. Every vector renders a certain distribution of activity within the connectionist network. The syntactic operations of the language have tensor operations as counterparts. Barnden (1991) pursues a related approach. As far as syntax is concerned, some connectionist networks can completely implement compositional languages.

The kind of compositionality that is necessary for systematicity, however, focuses not on syntactic, but on semantic features. The capacity to think that a child with a red coat is distracted by an old herring is not correlated with the capacity to think that a child with an old coat is distracted by a red herring. The thoughts ought to be correlated, though, if syntactic composition was sufficient for systematicity. Although both thoughts are syntactically composed from exactly the same primitives by exactly the same operations, they are not correlated because red herring is idiomatic, i.e. because the mapping (red, herring) $\rightarrow$ red herring is syntactically, but not semantically compositional. One may well have the capacity to think of red coats and old herrings even though one lacks the capacity to think of red herrings. We may infer, thus, that semantic compositionality is necessary for systematicity and that syntactic compositionality is not sufficient. The strategy to map the syntax of a systematic language homomorphically onto a connectionist network does not suffice to establish that the network itself is systematic.

To put the dilemma in a nutshell, connectionist models seem to be too weak to explain systematicity, whereas classical models are apparently too strong to be implemented by real brains. The rest of the paper will explore the option of something "in between" classical and connectionist architectures. The presented solution differs significantly from other approaches to the dilemma (Lange & Dyer, 1989; Shastri & Ajjanagadde, 1993; Hummel & Holyoak, 1997; Sougné 1999). Especially with respect to the representation of relations, the

presented model might have more plausible implications.

Constituency

A further argument provides us with a deeper insight into what's wrong with traditional connectionist networks and gives us a key how to match this deficit. Most semantic theories explain the semantic properties of internal representations either in terms of covariance, in terms of inferential relations, or in terms of associations. One may, e.g., hold that a certain internal state is a representation of redness because the state covaries with nearby instances of redness. This covariance relation is, of course, backed by the intrinsic and extrinsic causal properties of the redness representation. One may also hold that bachelor representations characteristically are such that the subject is disposed to infer unmarried-man representations from it. Those dispositions, again, are grounded in the causal properties of bachelor and unmarried-man representations. One may, thirdly, hold that the semantic value of the cow representation is determined by the fact that it is associated with representations like milk, leather, mammal, grass etc. The mechanism of association, too, supervenes on the causal properties of the representations in question. All of these theories have one principle in common: An internal representation has its semantic value because it has a certain causal role within the system (and – perhaps – the rest of the world). The question of how the semantic values of primitive representations determine the semantic value of complex representations, hence, leads to the question of how causal properties can be inherited from primitive to complex states. From chemistry we know that atoms determine the causal properties of molecules because atoms are constituents of molecules. Physics gives similar answers with regard to atoms and elementary particles. One can even make it a hard metaphysical point: If the causal properties of a state B are determined by, and dependent on, the causal properties of the states A_1, ..., A_n and their relations to each other, then A_1, ..., A_n are constituents of B. Here, constituents are conceived of as necessary parts: A is a constituent of B if and only if the following is necessary and generally true: If B occurs at a certain region of space at a certain time, then A occurs at the same region at the same time.

The failure of connectionist attempts, therefore, is that the homomorphism between language and network structure does not preserve the constituent relations within the language. The network counterparts of brown and cow aren't constituents of the network counterpart of brown cow. Since the homomorphism does not preserve constituent relations, it fails to transfer semantic compositionality: Although the operation (brown, cow) → brown cow is semantically compositional, the network operation (h(brown),h(cow)) → h(brown cow) may not be semantically compositional (h being the homomorphism). If h(brown) and h(cow) aren't constituents of h(brown cow) you cannot say: h(brown cow) co-varies with brown cows because h(brown) co-varies with brown things and h(cow) covaries with cows. If the constituent relations were preserved, you could say this. For the same reason, you are deprived of the possibility to explain the inferential and the associative properties of the complex representation on the basis of the inferential and the associative properties of the primitive representations.

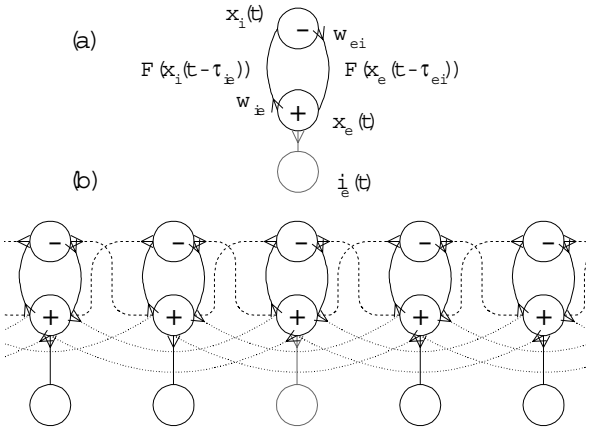


Figure 1: (a) Elementary oscillator consisting of a coupled pair of an excitatory (+) and an inhibitory unit (–) together with an input unit. t, time; x(t), unit activity; F(x) sigmoidal output function; w, coupling weight; τ, delay time; $i_e(t)$, external input. Subscripts: e, excitatory unit; i, inhibitory unit. (b) Oscillatory elements coupled by short-range synchronizing connections (dashed) and long-range desynchronizing connections (dotted), without interaction at crossings. The figure is meant to show the principle of coupled oscillators, rather than a particular connectivity pattern.

Synchrony

Constituency is a synchronic relation, while causal connectedness is a diachronic relation. Whole and part co-exist in time, whereas causes and effects succeed in time. The reference to causal connections and the flow of activation within the network will, therefore, not suffice to establish constituent relations. What we, in addition, need is an adequate synchronic relation. Oscillatory networks provide a framework to define such a relation: the relation of synchrony between phases of activity. Synchrony and asynchrony are synchronic relations because the relata, phases of activity, coexist in time. An elementary oscillator is realized by coupling an excitatory unit with an inhibitory unit using delay connections. An additional unit allows for external input (figure 1a). Within the network, oscillatory elements are coupled by either short-range synchronizing connections or long-range desynchronizing connections

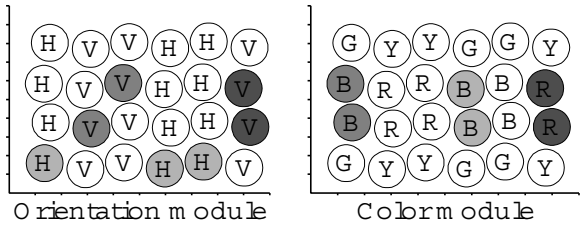


Figure 2: Scheme of a typical neuronal response aroused by a blue vertical, a red vertical, and a blue horizontal object. Circles with letters signify neurons with the property they indicate (V, H: horizontal, vertical; R, G, B, Y: red, green, blue, yellow). Like shadings signify synchronous activity. The phases of some blue-neurons are synchronous with the phases of some vertical-neurons (middle-shading), some phases of vertical-neurons are in synchrony with some phases of red-neurons (dark-shading), and some blue-neurons fire in phase with some horizontal-neurons (light-shading).

(figure 1b). A multitude of oscillators can be arranged in feature modules (e.g., the color module), employing appropriate patterns of connectivity. Given a certain selectivity of the input unit, each oscillator is designed to indicate a certain property (e.g., redness) within the feature domain. Oscillators for like properties are connected synchronizingly; those for unlike properties are connected desynchronizingly. The pattern of connectivity may as well reflect topographical features. The behavior of oscillatory networks have been studied in detail elsewhere (Schillen & König, 1994). Stimulated oscillatory networks, characteristically, show object-specific patterns of synchronized and desynchronized oscillators within and across feature modules. Oscillators that represent properties of the same object synchronize because oscillatory networks implement the Gestalt principles. Oscillators that represent properties of different objects desynchronize. We observe that for each represented object a certain phase of activity spreads through the networks. The phase pertains only to oscillators that represent the properties of the object in question. Assuming that elementary oscillators are models of neurons and that oscillatory networks are models of part of the visual cortex, the results of these studies support two hypotheses:

Indicativity. As part of the visual cortex, there are collections of neurons whose function it is to show activity only when an object in the perceptual field instantiates a certain property.

Synchrony. Neurons that belong to two collections indicative for the properties π_1 and π_2, respectively, have the function to show activity synchronous with each other only if the properties π_1 and π_2 are instantiated by the same object in the perceptual field.

The hypotheses are supported by neurobiological evidence. The indicative function of neurons was discovered by Hubel and Wiesel (1962, 1968). Neurobiologists meanwhile have specified a great variety of feature domains: color, orientation, direction of motion, speed, luminance, etc. Property indicative neuronal collections will, subsequently, be called π-collections, with π standing for the property the neurons of the collection indicate.

A number of experimental data support the hypothesis of synchrony (detailed review: Singer & Gray, 1995).[1] Synchrony of neurons (<200µm apart) within one column was recorded in many different species and cortical regions of awake and lightly anaesthetized animals, and can be observed in the local field potential as well as in the multi-unit and paired single-unit recordings (Gray & Singer, 1989; Kreiter & Singer, 1992). Intercolumnar synchrony of distant neurons (>2mm) was shown by simultaneously recording the activity of neurons in different parts of the cortex (Schwartz & Bolz, 1991). Its occurrence within and between visual areas depends upon whether the neurons are stimulated by single or separate objects. For example, synchrony is strong when two neurons in V1 with non-overlapping but collinear preferred orientations are stimulated by a single long bar moving across their receptive fields (Gray et al., 1989). It is weaker when they are stimulated by two short collinear bars moving in the same direction, and it is absent altogether when the two short bars move in opposite directions. These and other results support the view that the synchrony of distributed activity in the visual system implements the well-established Gestalt principles of perceptual grouping. The issue of object-binding as stated by the principle of synchrony is supported by evidence from the primary visual cortex of the cat (Engel, König, & Singer, 1991) and other animals. These experiments show that when two neurons with different orientation and direction preferences are stimulated by a single moving bar that is sub-optimal for both, then they synchronize, but when they are stimulated by two separate bars, each being optimal for one of the neurons, then they do not. The representational function of synchrony is supported by studies of binocular rivalry with awake strabismatic cats (Fries et al., 1997). There has long been anatomical evidence for long-range horizontal connections in V1 (Rockland & Lund, 1983). Lowel and Singer (1992) observed that these connections play a synchronizing role. Figure 2 provides a schematic overview.

Algebra

Oscillatory networks that implement the two hypotheses can be given an abstract algebraic description:

$$N = \langle N_i, N_p, N_s; \varphi_1, \ldots, \varphi_m; F_1, \ldots, F_n; \approx, \neq, \varepsilon, \wedge \rangle.$$

[1] O'Keefe and Recce (1993), Wehr and Laurent (1996), Gawne, Kjaer and Richmond (1996) assume a more critical attitude with respect to the role of synchrony in object binding.

Below, this algebra will be shown to be isomorphic to a systematic language. The primitive entities of the algebra are (i) the phases of activity picked out by the symbols φ_1, ..., φ_m and (ii) the sets of phases related to each π-collection and referred to by the symbols F_1, ..., F_n. The phases of activity are elements of the set of all neuronally possible phases N_i. The sets of phases are elements of N_p. The operations denoted by the symbols $\approx$, $\not\approx$, ε, and $\wedge$ serve to build complex neuronal states from primitive entities. The set of all complex neuronal states constructible in N is N_s. Superior "N" signifies that symbols or sequences of symbols in square brackets are interpreted in the algebra N. Thus φ_1^N, φ_2^N, ..., φ_m^N are phases of activity; F_1^N, F_2^N, ..., F_n^N are sets that comprise the phases of related π-collections; and $\approx^N$, $\not\approx^N$, ε^N, $\wedge^N$ are operations. Instead of F_1, F_2, ..., we will sometimes use more suggestive capital letters like the H, V, R, G, B, and Y of figure 2.

In N there is only one fundamental operation: being synchronous with. It is referred to by the operation symbol $\approx$ and relates phases of activity to each other:

$[\varphi_i \approx \varphi_j]^N$ is the state $[\varphi_i$ is synchronous with $\varphi_j]^N$.

The remaining N-operations are derivationally defined by means of standard symbols, with "¬" and "&" signifying negation and conjunction, "∃" the existential quantifier, "x" a variable, "(" and ")" prentices, "∈" set membership. We can thus define asynchrony $\not\approx^N$ in a natural way:

$[\varphi_i \not\approx \varphi_j]^N$ is the state $[\neg\, \varphi_i \approx \varphi_j]^N$.

If neurons of a π-collection, to which the set of phases F_j^N is assigned, show a certain phase of activity φ_i^N, we say that the phase φ_i^N or a synchronous equivalent is an element of the set F_j^N. To refer to this neuronal state, we define the relation of pertaining ε^N:

$[\varphi_i \varepsilon F_j]^N$ is the state $[(\exists x)(x \approx \varphi_i \,\&\, x \in F_j)]^N$.

A further operation is co-occurrence $\wedge^N$ of two states p^N and q^N. It is trivially defined:

$[p \wedge q]^N$ is the state $[p \,\&\, q]^N$.

The four operations are motivated by the hypothesis of indicativity and synchrony. They allow us to give an algebraic description of the scheme shown in figure 2. Assuming that the middle-shaded neurons show the phase of activity φ_1^N, the dark-shaded neurons the phase φ_2^N and the light-shaded neurons the phase φ_3^N, figure 2 expresses the cortical state:

$[\varphi_1 \varepsilon V \wedge \varphi_1 \varepsilon B \wedge \varphi_2 \varepsilon V \wedge \varphi_2 \varepsilon R \wedge \varphi_3 \varepsilon H \wedge \varphi_3 \varepsilon B]^N$.

Language

The notation already suggests that the algebra N might be isomorphic to a compositional and systematic language L. Since languages can be treated as algebras, we may define:

$L = \langle L_i, L_p, L_s; \varphi_1, ..., \varphi_m; F_1, ..., F_n; \approx, \not\approx, \varepsilon, \wedge \rangle$.

The entities of L are indexical expressions like this and that (included in the set L_i), predicates like red and vertical (in L_p) and clauses like this is red or this is the same as that (in L_s). The primitive symbols φ_1, ..., φ_m pick out specific indexicals and the primitive symbols F_1, ..., F_n specific predicates. Again we will sometimes use more suggestive capital letters instead of F_1, ..., F_n. The fundamental operation of L is sameness $\approx^L$:

$[\varphi_i \approx \varphi_j]^L$ is the clause $[\varphi_i$ is the same as $\varphi_j]^L$.

The remaining operations can derivationally be defined. Difference $\not\approx^L$:

$[\varphi_i \not\approx \varphi_j]^L$ is the clause $[\neg\, \varphi_i \approx \varphi_j]^L$.

Using $\in$ as the symbol for predication this time, the copula ε^L, which links an indexical expression φ_i^L to a predicate F_j^L, is defined by:

$[\varphi_i \varepsilon F_j]^L$ is the clause $[(\exists x)(x \approx \varphi_i \,\&\, x \in F_j)]^L$.

The copula (English: "is") enables us to paraphrase natural language sentences like this is vertical in L: $[\varphi_1 \varepsilon V]^L$. The conjunction $\wedge^L$ between two clauses p^L and q^L is defined:

$[p \wedge q]^L$ is the clause $[p \,\&\, q]^L$.

The sentence there is a blue vertical, a red vertical, and a blue horizontal object can now be paraphrased:

$[\varphi_1 \varepsilon V \wedge \varphi_1 \varepsilon B \wedge \varphi_2 \varepsilon V \wedge \varphi_2 \varepsilon R \wedge \varphi_3 \varepsilon H \wedge \varphi_3 \varepsilon B]^L$.

Isomorphism and Preserved Constituency

To prove that the algebras N and L are isomorphic, a number of conditions have to be warranted. (i) There are as many phases of activity in N as there are indexical terms in L. (ii) Each π-collection, respectively, each related set of phases in N is assigned to exactly one predicate of L. (iii) L-clauses, by stipulation, are identical if and only if they are logically equivalent. For, cortical states are identical if and only if they are referred to by logically equivalent N-descriptions. To ensure this non-trivial condition, we thus have to accept that order is irrelevant in L. This leads to a non-standard notion of language: Concatenation, no longer, is the fundamental operation of concept composition. (iv) The two fundamental operations synchrony and sameness are isomorphic. If so, this isomorphism then conveys to all operations that have recursively been defined. Since sameness is a reflexive, symmetric, and transitive relation, we have to define synchrony between phases as a reflexive, symmetric, and transitive relation, too. This is consistent with recent neurobiological data (cf. Eckhorn, 2000) and the computer simulations of oscillatory networks mentioned above.

In previous sections we argued that an architecture might not be compositional even if it is syntactically homomorphic to a compositional language. To preserve semantic compositionality, the isomorphism between L and N must, in addition, preserve the constituent structure of the language. If a primitive term is a constituent of a complex term, the isomorphic counterpart of the primitive term must be a constituent of the isomorphic counterpart of the complex term. The primitives of L are the indexicals φ_1^L, ..., φ_m^L and the predicates F_1^L, ..., F_n^L. Every L-operation will lead to targets with those primitives as constituents. The clause $[\varphi_1 \approx$

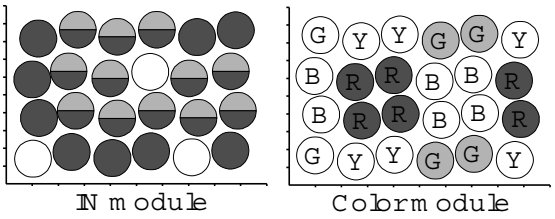


Figure 3: Predicted neuronal representation of relations. The state $[\varphi_1 \varepsilon G \wedge \varphi_2 \varepsilon R \wedge \langle \varphi_1, \varphi_2 \rangle \varepsilon' IN]^N$ is shown. The phases φ_1 of the G-neurons (light shading) occurs on the IN-module only in superposition with the phase φ_2 of the R-neurons (dark-shading) forming the duplex phase $\{\varphi_1, \varphi_2\}$ (hybrid shading). Since φ_2 also occurs as simplex on the IN-module, the situation on the IN-module is rendered by $[\{\{\varphi_1, \varphi_2\}, \{\varphi_2\}\} \varepsilon' IN]^N$. By definition, this is equivalent to $[\langle \varphi_1, \varphi_2 \rangle \varepsilon' IN]^N$.

$\varphi_2]^L$ can impossibly be tokened without tokening the indexicals φ_1^L and φ_2^L. With respect to constituency, what is true for L is also true for N: The state $[\varphi_1 \approx \varphi_2]^N$ is tokened just in case the phases φ_1^N and φ_2^N are tokened. Two phases are synchronous only if both of them actually occur within the cortex. The same is true mutatis mutandis for asynchrony. In L, the indexical φ^L and the predicate F^L are constituents of the clause $[\varphi \varepsilon F]^L$. Therefore, the phase φ^N and the π-collection to which the set F^N relates must be tokened, whenever the cortex is in $[\varphi \varepsilon F]^N$. This is obviously true because φ^N cannot pertain to the π-collection unless both the phase and the π-collection occur in the cortex. Figure 2 illustrates that the isomorphism preserves constituent relations for every operation: The complex state shown can only be tokened if, indeed, certain bursts of activity and certain collections of neurons are tokened. We may infer that oscillatory networks are not only syntactically, but also semantically compositional.

Relations

The representation of relations poses a binding problem of second order. The sentence this red vertical object is in that green horizontal object not only binds four property representations into two object representations, it moreover binds the two object representations by the relation in. The constituency preserving isomorphism between L and N straightforwardly generates a prediction of how to realize relational representation by oscillatory networks: After L has been extended by the tools for representing relations known from logic, N has to be extended in a way that perpetuates the isomorphism and the congruence with respect to constituency structure. The tools needed in the extensions of L and N are the operation of pairing, a higher-order copula and relation constants, or, respectively, their neuronal counterparts. Following Kuratowski (1967), ordered pairs are by common standards defined as asymmetric sets of second order:

$$[\langle \varphi_i, \varphi_j \rangle]^{L/N} =_{def} [\{\{\varphi_i, \varphi_j\}, \{\varphi_j\}\}]^{L/N}.$$

With the relations R_1^L, ..., R_k^L being sets of pairs, the higher-order copula links pairs to relations in the manner of set membership. On the neuronal level, the R_1^N, ..., R_k^N can be interpreted as relational modules:

$$[\langle \varphi_i, \varphi_j \rangle \varepsilon' R_1]^{L/N} =_{def} [\langle \varphi_i, \varphi_j \rangle \in R_1]^{L/N}.$$

The sentence this green object is in that red object can now be paraphrased in the extension of L:

$$[\varphi_1 \varepsilon G \wedge \varphi_2 \varepsilon R \wedge \langle \varphi_1, \varphi_2 \rangle \varepsilon' IN]^L.$$

Its neuronal counterpart – superior "L" is replaced by superior "N" – is shown in figure 3. To achieve a distribution of phases thus complex, some neurons are required to show a superposition of two phases. The presented model, therefore, predicts multiplex activity as a means of representing relations. Gasser and Colunga's (1998) simulation, which also uses superposed phases in relational representations, supports the prediction.

Neither Connectionism nor Classicism

Cognitive architectures can be distinguished along three features:

Syntactic Trees. There are mappings from ordered sets of argument representations onto target representations.

Constituency (presupposes trees). For every syntactic tree, its argument representations are constituents of its target representation.

Order (presupposes constituency). For every target representation, there is a determinate order among its constituents.

These features are each realized by every standard language: There is a syntax, words are constituents of phrases, and the words follow a determinate word order. We can now ask which of these features a certain cognitive model implements. Turing-style computers typically implement all three features because they build complex representations from primitive representation by concatenation following certain syntactic rules. Integrated connectionist/symbolic architectures only implement syntactic trees. They do not implement the principle of constituency and the principle of order. Oscillatory networks, however, implement both syntactic trees and the principle of constituency. They do not implement an order among representations.

Oscillatory networks lie in some sense in between classical and connectionist architectures. They resemble connectionist networks in many respects: They may serve as associative, content addressable memories. They process information in parallel. They are able to learn from examples. They degrade gracefully. Etc. Still, oscillatory networks are stronger than traditional connectionist networks because, in oscillatory networks, primitive representations are constituents of complex representations. The primitive representations

inherit their causal properties to complex representations and, thereby, determine their semantic properties. Oscillatory networks unite the virtues and avoid the vices of classical and connectionist networks. They are semantically compositional and systematic.

Acknowledgments

Research for this paper was sponsored by the National German Scholarship Foundation. It was enabled by a one-year research scholarship at Rutgers University and the Rutgers Center of Cognitive Science. I owe many of the presented insights to seminars and discussions with Andreas Engel, Thomas Metzinger, Wolf Singer, Jerry Fodor, Ernie LePore, Brian McLaughlin, Bruce Tesar and Gerhard Schurz. I am grateful, also, to the Berlin Colloquium "Philosophy Meets Cognitive Science".

References

Barnden, J.A. (1991). Encoding complex symbolic data structures with some unusual connectionist techniques. In J.A. Barnden & Pollack, J.B. (eds.). Advances in Connectionist and Neural Computation Theory Vol. 1: High-level connectionist models. Norwood, NJ: Ablex Publishing Corp.

Eckhorn, R. (2000). Neural mechanism of scene segmentation (abstract). Symposium on Neural binding of space and time. Leipzig: unpublished.

Engel, A.K., Konig, P, & Singer, W. (1991). Direct physiological evidence for scene segmentation by temporal coding. Proceedings of the National Academy of Sciences, USA, 88, 9136-40.

Fodor, J.A. & Pylyshyn, Z.W. (1988). Connectionism and cognitive architecture: A critical analysis. Cognition, 28, 3-71.

Fries, P., Roelfsema, P.R., Engel, A.K., König, P., & Singer, W. (1997). Synchronization of oscillatory responses in visual cortex correlates with perception in interocular rivalry. Proceedings of the National Academy of Sciences, US, 94, 12699–12704.

Gasser, M. & Colunga, E. (1998). Where Do Relations Come From? (Tech. Rep. 221). Bloomington, IN: Indiana University Cognitive Science Program.

Gawne, T.J., Kjaer, T.W., & Richmond, B, J. (1996). Latency: Another potential code for feature binding in striate cortex. Journal of Neurophysiology, 76, 1356-1360.

Gigerenzer, G. & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating, and perspective change. Cognition, 43, 127-171.

Gray, C.M. & Singer, W. (1989). Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. Proceedings of the National Academy of Sciences, USA, 86, 1698-702.

Gray, C.M., Konig, P., Engel, A.K. & Singer, W. (1989). Oscilliatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. Nature, 338, 334-37.

Horgan, T. & Tienson, J. (1996). Connectionism and the Philosophy of Psychology. Cambridge, MA: The MIT Press.

Hubel, D.H. & Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. Journal of Physiology, 160, 106-154.

Hubel, D.H. & Wiesel, T.N. (1968). Receptive fields and functional architecture of monkey striate cortex. Journal of Physiology, 195, 215-243.

Hummel, J.E. & Holyoak, K.J. (1997). Distributed representation of structure: A theory of analogical access and mapping. Psychological Review, 104, 427 – 466.

Kreiter, A.K. & Singer W. (1992). Oscillatory neuronal responses in the visual cortex of the awake macaque monkey. European Journal of Neuroscience, 4, 369-75.

Kuratowski, K. (1967). Set Theory. Amsterdam: North-Holand.

Lange, T.E. & Dyer, M.G. (1989). High-level inferencing in a connectionist network. Connection Science, 1, 181 –217.

Lowel, S., & Singer, W. (1992). Selection of intrinsic horizontal connections in the visual cortex by correlated neuronal activity. Science, 255, 209-12.

O'Keefe, J. & Recce, M. (1993). Phase relationship between hippocampal place units and the hippocampal theta rhythm. Hippocampus, 3, 317-330.

Rockland, K. & Lund, J.S. (1983). Intrinsic laminar lattice connections in primate visual cortex. Journal of Comparative Neurology, 216, 303-18.

Schillen, T.B. & König, P. (1994). Binding by temporal structure in multiple feature domains of an oscillatory neuronal network. Biological Cybernetics, 70, 397-405.

Schwartz, C. & Bolz, J. (1991). Functional specificity of the long-range horizontal connections in cat visual cortex: a cross-correlation study. Journal of Neuroscience, 11, 2995-3007.

Shastri, L. & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables, and dynamic bindings using temporal synchrony. Behavioral and Brain Sciences, 16, 417-94.

Singer, W. & Gray, C.M. (1995). Visual feature integration and the temporal correlation hypothesis. Annual Review of Neuroscience, 18, 555-86.

Smolensky, P. (1995). Connectionism, constituency and the language of thought. In Macdonald, C., & Macdonald, G. (Eds.), Connectionism. Cambridge, MA: Blackwell.

Sougné, J.P. (1999). INFERNET: A Neurocomputational Model of Binding and Inference. Doctoral Thesis, Université de Liège.

Wehr, M. & Laurent, G. (1996). Odour encoding by temporal sequences of firing in oscillating neuronal assemblies. Nature, 284, 163-66.

A Model of Perceptual Change by Domain Integration

Gert Westermann (**gert@csl.sony.fr**)
Sony Computer Science Laboratory
6 rue Amyot
75005 Paris, France

Abstract

A neural network model is presented that shows how the perception of stimuli can be changed due to their integration between different domains. The model develops representations based on the correlational structure between the stimuli of the domains. It shows how prototypicality, categorical perception, and interference effects between the domains can arise without the explicit labeling of categories. The model is extended to learn the sensori-motor integration between articulatory parameters and speech sounds and it is shown how it can, in accordance with the ideomotor principle, imitate sounds based on the developed mappings in a "babbling phase".

Introduction

The ability to categorize is one of the most fundamental cognitive processes. Nevertheless, uncovering the mechanisms that underlie this ability has challenged experimenters and modelers alike. The reason for this difficulty might be that categories can be formed in many different ways: in some cases, perhaps mainly in experimental situations, explicit information about the category of a stimulus is given. In other cases, no feedback might be available about the categorization choice, and in even others, no explicit categorization choice might be made at all.

At the same time, recent research has suggested that categorization itself can exert an influence on perception (Goldstone, 1995; Schyns *et al.*, 1998). While these effects have mainly been studied in a supervised paradigm, perceptual changes also occur prominently in categorization without supervision and without explicit labeling, for example, in being exposed to the phonemes of one's native language (Kuhl *et al.*, 1992).

Finally, there is clear evidence that in categorizing the world, we make use of all available information and integrate the information from different modalities, making categorization more robust and easier. For example, visual and auditory information are integrated in speech perception, leading to enhanced activity in the cortical areas responsible for both domains (e.g. Calvert *et al.*, 1999).

In this paper a neural network model is described that aims to integrate several aspects of categorization, namely, the combination of modalities and the perceptual changes that go hand in hand with categorization. The model suggests that some of the phenomena that are usually explained as the consequence of explicit categorization, e.g., prototype formation and categorical perception, can arise without such explicit categorization based on the correlational structure of the stimuli from different modalities, and that they can facilitate subsequent explicit categorization when it occurs.

The integration between modalities has previously been modeled by de Sa and Ballard (1998). Their neural network model consisted of one layer for each modality, and each layer made an explicit category decision. In a process of self-supervised learning both modalities learned to agree on their decision. While the model performed comparably to supervised models, it was necessary to determine the number of categories *a priori*, and due to absolute category boundaries in each modality, perceptual phenomena such as gradedness and varying sensitivities to differences between stimuli could not be modeled. The present model aims to give an account of how such perceptual phenomena that are usually linked with explicit categorization can occur without an explicit category decision and without labeling, and how such a model can be extended to also account for sensori-motor integration. The model is loosely inspired by neurobiological considerations.

The rest of the paper is organized as follows: first, the model is described in detail. Then, experiments with a simple data set are described that lead to perceptual change as the result of the integration between modalities. Finally, the application of the model to sensori-motor integration and the imitation of sounds is described.

The Domain-Integration Model

The model described here integrates the stimuli from two domains (modalities) into a unified percept. The architecture of the model is shown in fig. 1. Each domain is represented by a neural map, and Hebbian connections between the maps allow for the coordination between them. Usually, an input pair (one input per map) is presented to the maps simultaneously, and in the following the activation and weight update mechanisms are described.

Each neural map consists of a number n of units that are randomly positioned in the input space (in this paper, the input spaces for both domains are two-dimensional

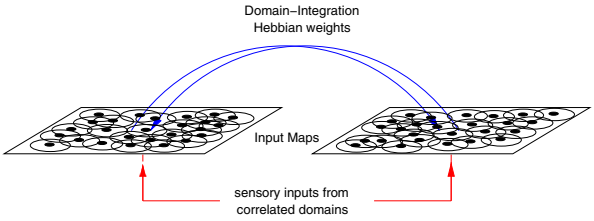


Figure 1: The architecture of the model.

to facilitate visualization). In the current model, the positions of these units remain fixed throughout learning. Each unit acts as a receptive field with a Gaussian activation function of a fixed width. Such receptive fields exist in many areas of the cortex. When an external input x is presented to the map, the Gaussian activation of the units is computed as

$$act_{i_{ext}} = e^{\frac{pos_i - x}{\sigma^2}} \tag{1}$$

where pos_i is the position of unit i, x is the input signal, and σ is the standard deviation (width) of the Gaussian. Each unit is connected with unidirectional Hebbian weights to all units on the map for the other domain. The Hebbian activation of a unit is the dot product of the weight value vector and the activation vector of the units on the other map:

$$act_{i_{hebb}} = \sum_k act_k w_{ik} \tag{2}$$

where the units on the other map are designated by the index k and w_{ik} is the weight from a unit k on the other map to the current unit i on this map.

The total activation of a unit is computed by summing the activations from the external stimulus and those from the Hebbian connections with the other map:

$$act_i = \gamma_e act_{i_{ext}} \quad \gamma_h act_{i_{hebb}} \tag{3}$$

where γ_e and γ_h are weighting parameters to control how much each partial activation contributes to the total activation of the unit.

The activation update after presentation of a pattern is synchronous for all units, and the activation values are scaled to a maximum of 1.0.

One input to a map will typically activate several units, and the response r_i to an input x, that is, how the neural map "perceives" that input, is computed by a population code: the response is the vector sum of all units i, weighted by their activation values:

$$\mathbf{r}_x = \frac{\sum_i act_i \mathbf{pos}_i}{\sum_i act_i} \tag{4}$$

Such population codes have been found to play a role for example in the encoding of motor commands in the monkey cortex (Georgopoulos *et al.*, 1988) where the direction of arm reaching is predicted accurately by adding the direction vectors of direction sensitive neurons, weighted by their firing rate. In computational models, population codes have been successfully used to show the emergence of a perceptual magnet effect for phonemes (Guenther and Gjaja, 1996).

The Hebbian connections between the maps are updated with the covariance learning rule (Sejnowski, 1977):

$$\Delta w_{ik} = \alpha(act_i - \bar{act}_i)(act_k - \bar{act}_k) \tag{5}$$

where $\bar{act}_i$ and $\bar{act}_k$ are the average activation values of units i and k over a certain time interval. This rule strengthens the connections between units when their activation values are positively correlated, weakens them when the activations are negatively correlated, and does not change the weights when the activations are decorrelated.

This correlation-based weight update has the consequence that units that respond to stimuli that consistently co-vary across the domains develop higher activations due to the growing Hebbian weights: co-varying inputs in the two domains result in the same units on both maps to have co-varying activation values, and thus to develop strong Hebbian connections. This results in such units not only receiving external, but also strong Hebbian activation and becoming more active than other units that do not reliably co-vary with units from the other domain map. Given that the population code response is weighted by unit activations, this means that such units "pull" the response towards them and induce a perceptual change (fig. 2). Therefore, an input-pair with normal (previously observed) correlational structure will become more prototypical so that other, nearby inputs will be displaced towards it.

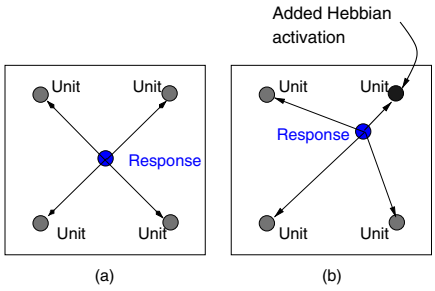


Figure 2: The response to an input is influenced by external Hebbian activation. (a): Without Hebbian activation, the response lies in the middle between four equally activated units. (b): When one unit is activated more due to Hebbian activation, the response is displaced towards that unit.

In the following section, experiments with this model are described that investigate the nature of the induced perceptual changes based on the integration between the two input domains.

Experiments

The domain-integration model was tested with a simple data set (fig. 3) to investigate the nature of the developed perceptual changes and the role of correlations between data from the two domains in this process.

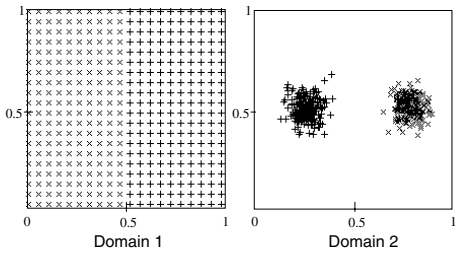


Figure 3: The data used to evaluate the model. The correlational structure between data items splits each domain into two classes, denoted by x and +, respectively.

Domain 1 consists of 400 evenly distributed two-dimensional data in the range from 0 to 1. Domain 2 consists of two clusters of 200 data each with Gaussian distributions around the centers $(0.25, 0.5)$ and $(0.75, 0.5)$. In training, the "left half" of the data in domain 1 (i.e., between 0.0 and 0.5) co-occurred with data from the "right" cluster of domain 2, and the "right half" in domain 1 (0.5 to 1.0) with data from the "left" cluster in domain 2.

Although this data set is artificial, it could be interpreted as, for example, a continuous variation of width and height of an object (domain 1) and associated sounds at certain frequencies and volumes (domain 2) in a modality-integration experiment.

The neural maps for each domain consisted of 200 randomly placed units. All data pairs were presented to the model a single time in randomized order. The Hebbian connections between the maps had initial values of 0 and were updated after presentation of each data pair. The parameter settings were $\alpha = 0.01$, and for each map, $\sigma = 0.05$, $\gamma_e = 1.0$, and $\gamma_h = 0.02$.

Development of Prototypes

Fig. 4 shows the initial and final responses to the data set. Each data input creates a response on its neural map (eq. (4)). Fig. 4A shows the initial response of the neural maps to the data from each domain. With all Hebbian connections being zero, the response is only determined by the actual input signal to the map and gives a rather faithful representation of the original data set in fig. 3. Due to the random location of units the original data is slightly distorted in the response.

During the training of the model, the Hebbian connections between units responding to co-varying data in both domains are strengthened and those responding to non-co-varying data are weakened or remain unchanged (eq. (5)). This process results in strong connections between units that respond to the centers of their categories because they will be active for both central and more

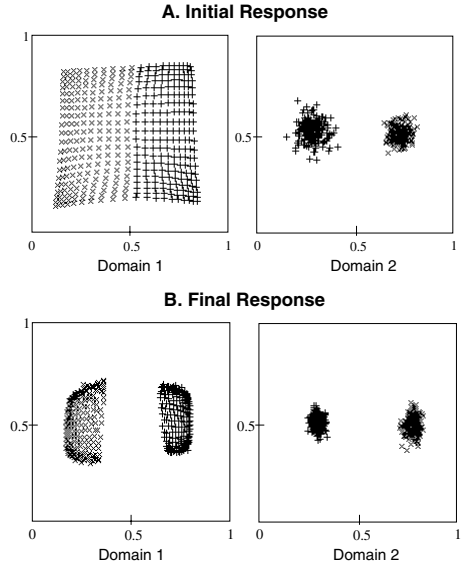


Figure 4: The initial (A.) and final (B.) response of the model to the data set in fig. 3.

peripheral inputs from a certain category. As a consequence, such central units will become more active than others when correlated inputs are presented. Their activation is a sum of the external activation caused by the inputs themselves, together with the activation mediated through the strengthened Hebbian weights from the other map (eq. (3)). Therefore, the response to peripheral stimuli will be pulled towards the center of each category. Fig. 4B shows the responses of the maps to the data after presentation of each data item and corresponding updating of the Hebbian connections. The continuous data in domain 1 has split into two clusters that correspond to the co-variance relations with the clusters in domain 2. Each cluster is based around a prototype determined by the central data item of each set. Similarly, the clusters in domain 2 have become very dense around their respective centers. Prototypes thus develop simultaneously in both domains, based on the interactions between the domain maps.

Categorical Perception

Categorical Perception (CP) is a phenomenon that occurs both innately and in learned categorization (see Harnad, 1987, for an overview): different stimuli within one category are perceived as more similar than stimuli with the same "distance" that straddle category boundaries. One example for innate CP is the perception of color, e.g. in a rainbow: although the light frequency changes continuously, we perceive separate bands of color. For example, within the red band we do not perceive differences between changing light frequencies, but an equally small change at the border of that band leads to the abrupt perception of orange.

It has been shown that CP also develops in learned categories such as phonemes of one's native language

(e.g. Kuhl *et al.*, 1992). More recently, CP has also been shown to arise for visual stimuli in categorization task experiments (Goldstone *et al.*, 1996). In these experiments, subjects had to group a set of continuously varied shapes into categories in a supervised learning task. After having learned the categories, they were better able to distinguish between two stimuli that were near a category boundary than between those that were within a category. Therefore, CP can be said to constitute a warping of similarity space in which the sensitivity to differences of (relevant) stimuli is enhanced near category boundaries and is decreased within categories.

Guenther and Gjaja (1996) modeled categorical perception for phonemes in an unsupervised model. They argued that the firing preferences of neurons in the auditory map reflect the distribution of sounds in the language, and due to the non-uniform distribution of these sounds CP arose in the model in a self-organizing process. While this model accounts well for CP in phoneme perception, it relies on a non-uniform distribution of the data. CP that arises for uniform stimuli as a result of explicit categorization has been modeled in supervised radial basis (Goldstone *et al.*, 1996) or backpropagation (Tijsseling and S.Harnad, 1997) networks. It therefore seems that CP can arise from different causes (data distribution or explicit teaching), and in the model presented here a third route is taken: it is studied how CP can arise in a homogeneously distributed data set that is correlated with non-uniform data in another domain, without the explicit teaching of category labels. Instead, categories form in an unsupervised way based on the correlational structures between the two domains.

In the present experiments, the x-coordinate of the data is the relevant dimension for determining category membership (with the categories defined by the correlations across domains). To establish whether CP did occur in the model, after training the map of domain 1 was presented with a sequence of data points from (0.0, 0.5) to (1.0, 0.5) in steps of 0.01, i.e., a walk from the left to the right side of the map. The difference between the responses of the model to every pair of adjacent data points is shown in fig. 5. There is a marked peak of sensitivity at the category boundary (0.5) where a difference of 0.01 in the input data is perceived as a difference of 0.08 in the responses. By contrast, at a distance from the category boundary, the sensitivity of the model to differences between stimuli is decreased.

This result models the basic characteristics of CP: an increased sensitivity to differences at the category boundary, and a diminished sensitivity within the categories.

Domain Integration: The McGurk Effect

Many experiments have shown that visual information can enhance the understanding of speech, suggesting an integration of the visual with the auditory signal in this task (see Massaro, 1998, for an overview). Striking evidence for the strength of this integration comes from the

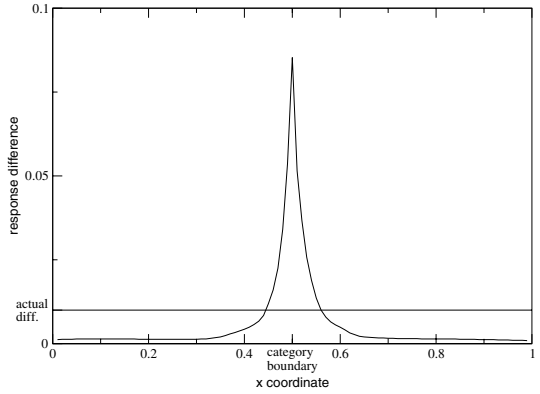


Figure 5: CP in the model: sensitivity to differences is increased around the category boundary.

McGurk effect (McGurk and MacDonald, 1976): when subjects are presented with conflicting visual and auditory data, their perception of what is said can be different from both the visual and the auditory signal. For example, when a face articulating /ga/ is dubbed with the auditory /da/, subjects perceive /ba/. This effect is highly pervasive and not subject to volitional control. It is not restricted to vision and auditory integration, but has also been found for touch and audition (Fowler and Dekle, 1991).

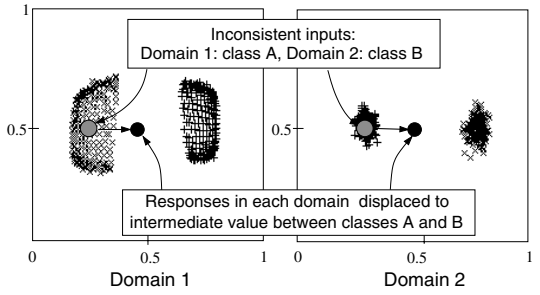


Figure 6: Exemplary response of the model to a data pair that does not correspond to the learned correlational structure. The previously learned responses are denoted by x and +, the data pair that does not correspond to the learned correlational structure by grey circles, and the response of the model to this data pair by black circles.

To test whether the model displayed a response similar to the McGurk effect in humans, data pairs were presented that did not correspond to the previously learned correlation structure. While during training the "left" half of the data set for domain 1 co-occurred with the "right" cluster in domain 2, now data from the "left" half in domain 1 was presented together with that from the "left" cluster of domain 2. Conceptually this corresponds to presenting e.g., an auditory /da/ together with a visual /ga/. The model integrated these conflicting inputs to a

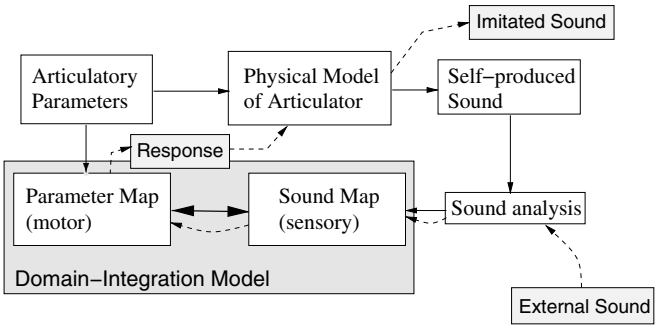


Figure 7: The model for sensori-motor integration and the imitation of sounds. Solid arrows indicate the "babbling phase" where the mapping from motor to sensory parameters is learned. The dashed arrows show the pathway for the subsequent imitation of sounds.

response that was a blend between the responses to each individual input (fig. 6).

While the McGurk effect has been studied in great detail and has revealed many results that are much more subtle than a simple blend between the auditory and visual information, the present model can give a principled account of how the domain integration that lies at the basis of this effect can arise. The details of the McGurk effect cannot be modeled with the artificial data set used here to investigate the general functionality of the model, but experiments are planned to use a more realistic set of auditory and visual signals that will give more detailed results.

In summary, the domain integration model displays, without the explicit teaching and labeling of categories, several of the effects that are generally supposed to rely on such labeling, namely, the formation of prototypes as attractors in the stimulus space, categorical perception in an evenly distributed set of stimuli, and an integration of stimuli from different domains to form a unified percept that forms a "compromise" when conflicting data is presented in the domains simultaneously.

The Model in Sensori-Motor Integration

In the previous sections it was described how the domain integration model integrates between two sensory domains, leading to psychologically observed phenomena such as prototype formation, categorical perception, and the McGurk effect. In this section, an extension to this model is proposed to account for sensori-motor integration (fig. 7). This extension works by presenting in one domain a representation of an action (e.g., motor parameters), and in the other, a representation of the sensory consequences of that action. The model then learns the associations between the motor commands and their sensory consequences, developing simultaneously in both domains prototypes of actions and consequences of these actions, based on a reliable correlation between them.

The sensori-motor variant of the model was tested on sound production. For this purpose, a physical model of a speech synthesizer (Boersma and Weenink, 1996) was used. In initial experiments, two parameters, jaw opening and the position of the styloglossus muscle (a muscle that controls the position of the back of the tongue) were varied continuously at 18 steps each, and the resulting sounds were analyzed with respect to their firsts two formant values. The model was trained on the resulting two-domain data set with 324 items. Fig. 8 shows the initial and final responses of the model. While the motor parameters are evenly distributed prior to training, after training prototypical parameter-sound pairs have formed in both domains due to their correlational structure.

The sensori-motor integration model corresponds to the ideomotor principle which postulates a tight coupling between perception and action. As such it can give an account of the imitation of sounds (fig. 7, fig. 8B): an external sound that is presented to the model evokes a response on the auditory map. This response is propagated through the developed Hebbian connections to the motor map where a motor response is evoked which can be used to articulate, i.e., imitate, the corresponding sound. However, the imitation of the heard sound is displaced towards a prototype that the model has developed during training (indicated by an arrow in the auditory map in fig. 8B). In this way, imitation is not merely a reproduction of an external stimulus, but a re-interpretation of that stimulus based on the developed structure of the model.

Discussion

The model described in this paper presents an algorithm to integrate sensory information between two domains to form a unified percept, thereby displaying phenomena also observed in human categorization. The model can be extended to also account for sensori-motor integration and the imitation of low level percepts. While the simple data sets used in this paper were used to demonstrate the principled functionality of the model, more realistic and

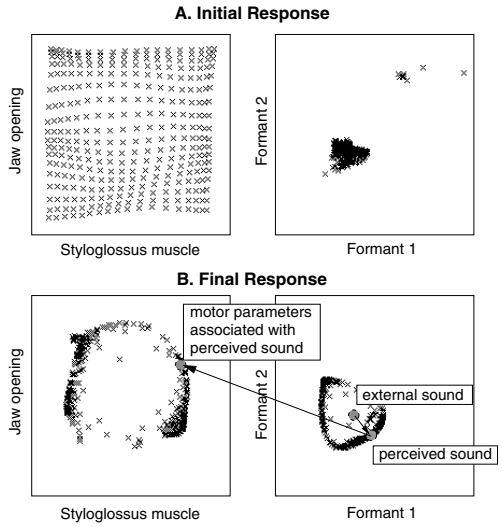


Figure 8: A. Initial and B. final responses of the sensori-motor integration model.

extensive experiments are necessary to establish whether it can account for more detailed results in these domains. We have now started to use higher-dimensional data for the learning of the articulation–perception mapping in sound production and imitation, and preliminary results look promising.

An important property of the model is that it shows a unified account of sensori-sensor and sensori-motor integration in a neurobiologically inspired framework.

An alternative view of this model could be as a variant of supervised category learning: when one map receives the inputs (i.e., object representations) and the other the targets (i.e., category labels), the model learns the mapping from the category members to their labels if there is a sufficient number of different categories. The domain integration model, however, adds an important aspect that is often neglected in supervised category learning models: not only category members, but also the concept of "category" has a topology and is changed by its members. For example, the "concepts" of the dog and cat categories will move closer together on the target map if their members share properties. In this way it becomes possible to measure the similarity between concepts by investigating the developed topology on the target map.

In its present form the model is simple, though it allows insights into how perception can change due to categorization. However, more realistic training data, as well as an extension of the model to be able to handle sequential and more complex data, are necessary. These will be the next steps in the described research.

Acknowledgments I would like to thank Eduardo Miranda for providing the data set for the sound imitation experiments.

References

Boersma, P. and Weenink, D. (1996). Praat, a system for doing phonetics by computer. Technical Report 132, Institute of Phonetic Sciences of the University of Amsterdam.

Calvert, G., Brammer, M., Bullmore, E., Campbell, R., Iversen, S., and David, A. (1999). Response amplification in sensory-specific cortices during crossmodal binding. *Neuroreport*, **10**, 2619–2623.

de Sa, V. R. and Ballard, D. H. (1998). Category learning through multimodality sensing. *Neural Computation*, **10**, 1097–1117.

Fowler, C. and Dekle, D. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, **17**, 816–828.

Georgopoulos, A. P., Kettner, R. E., and Schwartz, A. B. (1988). Primate motor cortex and free arm movements to visual targets in three-dimensional space. II. Coding of the direction of movement by a neural population. *Journal of Neuroscience*, **8**, 2928–2937.

Goldstone, R. L. (1995). Effects of categorization on color perception. *Psychological Science*, **6**, 298–304.

Goldstone, R. L., Steyvers, M., and Larimer, K. (1996). Categorical perception of novel dimensions. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, pages 243–248, Hillsdale, NJ. Erlbaum.

Guenther, F. H. and Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustic Society of America*, **100**, 1111–1121.

Harnad, S. (1987). *Categorical Perception*. Cambridge University Press, Cambridge.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, **255**, 606–608.

Massaro, D. W. (1998). *Perceiving Talking Faces*. MIT Press, Cambridge, MA.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746–748.

Schyns, P. G., Goldstone, R. L., and Thibaut, J. (1998). The development of features in object concepts. *Behavioral and Brain Sciences*, **21**, 1–54.

Sejnowski, T. J. (1977). Storing covariance with nonlinearly interacting neurons. *Journal of Mathematical Biology*, **4**, 303–312.

Tijsseling, A. and S.Harnad (1997). Warping similarity space in category learning by backprop nets. In M. Ramscar, U. Hahn, E. Cambouropolos, and H. Pain, editors, *Proceedings of SimCat 1997: Interdisciplinary Workshop on Similarity and Categorization*, pages 263 – 269. Department of Artificial Intelligence, Edinburgh University.

Imagery, Context Availability, Contextual Constraint and Abstractness

Katja Wiemer-Hastings (KATJA@NIU.EDU)
Jan Krug (JKDKRUG@YAHOO.COM)
Xu Xu (Z015504@STUDENTS.NIU.EDU)
Department of Psychology, Northern Illinois University, DeKalb, IL 60115, USA

Abstract

A constraint-based theory of abstractness was investigated according to which abstractness of entities is a function of (i) perceptual observability and (ii) characteristics of contextual constraints. Participants performed ratings of context availability, imagery, and abstractness for 36 nouns that varied in abstractness and familiarity. The ratings were used to compare the predictions of abstractness ratings by context availability, dual coding theory and the constraint-based approach outlined in this paper. We found that only constraints explain variation of perceived abstractness for abstract concepts, whereas context availability and imagery are good predictors of the dichotomous distinction of concrete-abstract, and of variations of concreteness for concrete concepts only. A second study shows that introspection-based constraints are most critical for abstractness ratings. Implications are discussed.

Abstractness

Every-day communication is pervaded by references to abstract entities, such as *explanation*, *regret*, and *intention*. Typically, we think of an entity as abstract when it cannot be perceived. However, there are no clear-cut criteria for what makes entities abstract or concrete. Several theoretical approaches exist to predicting perceived abstractness. This paper compares three theories: dual-coding theory, context-availability theory, and our approach, called the *contextual constraint theory*.

We propose that perceived abstractness depends on two factors. First, entities are abstract or concrete, depending on whether they are physical in nature (i.e., perceivable through vision, touch, etc.). Second, within these groups, abstractness varies according to more specific types of information. Together, we call this the *two-factor model of abstractness*. We will start by reviewing the plausibility of the dichotomy of abstract and concrete, as proposed by the first factor. The remainder of this paper will address the factors underlying abstractness variation within the groups of abstract versus concrete entities.

Abstract and Concrete: Dichotomy or Continuum?

Concrete and abstract nouns are commonly defined by reference to perceivability: Concrete entities are considered to be physical entities with characteristic shapes, parts, materials, etc., whereas abstract entities lack physical attributes (e.g., Crystal, 1995). The first proposed factor follows this broad distinction.

Some entities challenge the notion of a dichotomy of abstract and concrete entities. Examples for entities that cannot clearly be classified as abstract or concrete are *government*, *officer*, or *anger*. A *government* is abstract in that we cannot really point to who or what it is, but it is also concrete in that it involves a number of specific, concrete entities, such as people, buildings, and particular locations. *Officer* is a social agent term, referring to concrete individuals with characteristics defined by a particular social role or profession. Their roles are not obvious characteristics, but are inferred from more complex information, such as behavior patterns in specific situations.

Finally, emotion terms such as *anger* are a special group of entities. Emotions can be perceived within individuals who experience them. Outwardly, we can perceive emotion through nonverbal and verbal behavior. Still, emotions are qualitatively different from concrete entities such as cups and office chairs. In fact, they have been proposed to constitute a distinct group from both concrete and abstract entities (Altarriba, Bauer, & Benvenuto, 1999). The alternative view suggested by these challenges is a contiuum view, according to which all entities vary in concreteness, and the distinction of abstract versus concrete is an oversimplification.

A simple way to test both views is to ask people to rate the concreteness of a large sample of entities, including abstract and concrete ones. If concreteness is one dimension and all entities vary along this dimension, then ratings should be distributed pretty evenly across the entire scale. In contrast, if abstract and concrete entities were two distinct classes of entities, ratings should fall into two clusters. There would be a lot of entities rated as abstract, versus a lot of entities rated as concrete. That is, the distribution of

concreteness ratings would assume the shape of a bimodal distribution.

What is found is, in fact, that the ratings form two fairly distinct clusters with a different mode each. One mode is centered over the abstract half of the scale, the other mode is located over the middle of the concrete half. This finding has first been reported for 2172 words by Nelson and Schreiber (1992), and has been replicated here for an independently sampled set of 1660 nouns (see Figure 1).

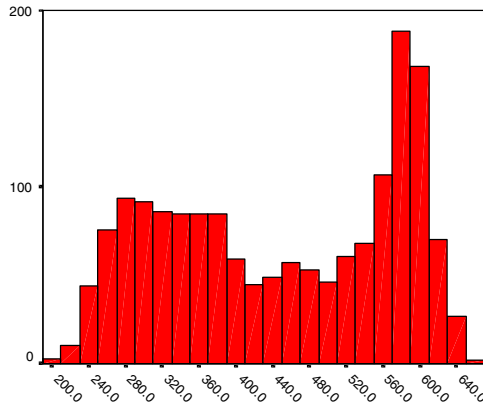


Figure 1
Distribution of concreteness ratings for 1660 nouns

The bimodal distribution is consistent with the view that abstract and concrete entities fall into two big clusters according to particular characteristics (e.g., tangibility or visibility). It is also obvious that, within these two clusters, entities vary in concreteness. What factors are causing this variance? The remainder of this paper addresses this question.

Variation in Abstractness

What information underlies the variance of abstractness in abstract entities? If an entity cannot be perceived, it is abstract. The lack of concreteness can account for it being abstract, but the same information cannot explain why some abstract entities are more abstract than others. For example, lack of perceptual information cannot explain why *principle* is rated more abstract than *idea*.

Our research aims to specify what factor(s) cause concrete entities to vary in perceived concreteness, and what factor(s) cause variance in the perceived abstractness of abstract concepts. Our studies are motivated by two lines of reasoning. The first directly follows from the *two-factor model of abstractness.* We do not think that perceivability accounts for the entire variation in concreteness. Instead, we assume a second factor. This intends to replace the commonsense notion that abstract concepts get more abstract to the degree that they get less perceivable (concrete). Second, it is quite likely that the difference between abstract and concrete is not so much a quantitative distinction, but a qualitative one. That is, it is conceivable that the factor or factors that make abstract entities more or less abstract are distinct from the factors that make concrete entities differ in concreteness. To identify possible factors, we next review theories that have been proposed to understand abstractness.

Theories of Abstractness

Theories pertinent to explaining abstractness include dual coding theory, context-availability theory, and contextual constraint theory. The predictions of each are discussed in turn.

Dual-Coding Theory / Imagery

The dual-coding model (Paivio, 1986) is one of the oldest theories about differences of abstract and concrete concepts. It proposes that the fundamental difference between abstract and concrete concepts is that only the concrete ones they are associated with imagery (henceforth IM) information, whereas both abstract and concrete concepts can be processed in a language-like code. The availability of two codes for the representation and processing of concrete concepts results in their processing advantage in many tasks, such as comprehension, word recognition and recall.

Applying the dual-coding theory to the prediction of perceived abstractness is fairly straightforward. The model is essentially dichotomous in its division of abstract and concrete, but one could derive the prediction from it that more concrete entities have higher imageability. If the dual-coding theory can be applied to variation of abstractness within the group of abstract entities, then one should find that abstract entities rated as most abstract are the entities that elicit the least imagery.

Context Availability Theory

The context availability theory (henceforth CA; Schwanenflugel & Shoben, 1983) argues that it is easier to think of a context for concrete objects than for abstract ones. The issue here is not whether a context can come to mind at all, but how long it takes to retrieve or construct it based on information in memory. Typical studies instruct participants to rate CA based on the time it takes to think of a context. If it takes a long time, then the rating should be low. If they can think of a context immediately, they should give a high rating.

Research has shown that CA ratings can account for a lot of effects labeled as concreteness effects, often even better than concreteness ratings themselves. If abstract words are more difficult to process because of less available context, then the prediction for abstractness ratings is that a word will be rated the more abstract, the less context is available for it.

Evidence Related to Context Availability

Rated CA has been shown to correlate highly with rated abstractness. Thus, it may offer a theoretical basis for predicting abstractness. However, Altarriba et al. (1999) found that the correlation of CA with concreteness differed for abstract, concrete, and emotion terms. Interestingly, the correlation was highest for concrete words (r = 0.68), second for emotion terms (r = 0.41), and lowest for abstract terms (r = 0.25). All three correlations were significant, but it is clear that the concreteness ratings for abstract words were only weakly related to CA, in comparison with the other groups.

The results cannot be applied directly to this work, because the sample used by Altarriba et al. was not limited to nouns. Another goal of the present study was to compare the correlations of ratings separately for abstract versus concrete entities, to examine whether the findings of Altarriba et al. hold up for a sample of nouns exclusively.

Contextual Constraint Theory

Abstract entities are associated with contexts (Schwanenflugel, 1991; Wiemer-Hastings & Graesser, 1998). They "apply" to, or are manifested in, situations. This application is contingent on particular events and circumstances in the situation. For example, an idea is contingent on an agent with a mental event, which will be expressed verbally or in some kind of behavior, and can be evaluated. An idea is thought of in one moment, expressed in another, maybe rejected in a third. As such, many abstract entities have characteristics akin to verbs: they are related to observable events in a situation, which are defined temporally.

Depending on the situation aspects that an abstract entity is contingent on, it can occur in many or few kinds of context. Roughly, the more particular situation elements are necessarily involved in its manifestations, the more constrained its occurrence is. An entity that is contingent only on few, and rather abstract situation characteristics (such as the presence of some entity) can occur in all kinds of situations.

Do contextual constraints affect the abstractness of abstract entities? Conceivably, an entity that is not strongly constrained is more abstract than an entity that is contingent on a fairly extensive set of constraints. Additionally, entities that only occur when concrete situation aspects are present may be less abstract than entities that are contingent on abstract, unobservable, or complex temporal elements of situations, or of information that is only accessible to introspection (such as a mental process). To test this contextual constraint theory, the materials for the study were coded for contextual constraints.

Constraints Naturally, situations vary. For example, *ideas* occur in various settings, through different agents, related to different problems, and varying in quality. However, the underlying constraints, such as the presence of an agent, remain largely unvaried. To examine the influence of such constraints, a list of contextual constraints was derived from a simplistic situation model, as exhaustive as possible. The list consists of agents, objects, "issues", mental states, relations, and temporal information. The constraints were selected to be relevant to abstract concepts (see Table 1). These constraints are not intended to describe situations with all the richness of information they contain, but to identify abstract building blocks of situations, without regard to their specific contents.

Table 1: Contextual constraints on abstract entities

Concrete elements	**Introspective elements**
Agent	Goal
Agent 2	Knowledge / memory
Group (people)	Belief / attitude
Object	Feeling
Location	Mental event / thought
Utterance	**Relations**
Action	Agent-agent
Object attribute	Agent-other people
Nonverbal behavior	Agent-object
Situation elements	Agent-thematic subject
Issue / topic	Relation between two entities
Obstacle	Utterance-issue relation
Temporal aspects	
Relevance of past	
Relevance of future	
Changes between time slices (Event)	
Continuity of change between time slices (Process)	
Continuity of state between time slices (State)	
Time-adjacency of events (causality)	

Such constraints can play a powerful role in the processing of abstract concepts. Assuming that context information must be accessed to comprehend the concept (e.g., Schwanenflugel, 1991; Schwanenflugel & Shoben, 1983), constraints can be used to guide the

mental construction of a context example. As such, they functionally resemble schemata and scripts (Schank & Abelson, 1977). Constraints fall into several groups, including concrete situation elements, object attributes, agent characteristics, situation elements, relations, and information about temporal characteristics and sequences.

Study 1: Comparing Accounts for Abstractness

The experiment systematically compared to what extent different theories can predict abstractness ratings. Participants were asked to make ratings of the predictor variables for a set of 36 words.

Materials

Words were randomly sampled from about 2000 nouns collected from the MRC2 database (Coltheart, 1981). An exhaustive search was made for nouns for which frequency estimates (Kucera & Francis, 1967), familiarity, and abstractness ratings (Gilhooly & Logie, 1980; Paivio, Yuille, & Madigan, 1968; Toglia & Battig, 1978) were available.

The sample was divided into 6 sets of different levels of abstractness, based on the abstractness ratings from the MRC2 database. The range was divided into six equal-sized parts, regardless of the number of words falling into each section. Words were matched in familiarity across groups to control for familiarity effects (see Kacinik, Shears, & Chiarello, 2000). From each of the groups, six words were randomly selected to be included in the study. The words were the concrete words *bass, beehive, blossom, hairpin, insect, labyrinth, lace, mackerel, morass, nectar, owl, pest, prize, sedative, tree, venom, vine,* and *zone* and the abstract words *aspect, day, daybreak, desperation, emancipation, exception, formation, happiness, hope, inaction, ingratitude, jeopardy, mischief, pity, possession, removal, saga,* and *story*.

Instructions

Instructions varied to elicit different kinds of information associated with the words presented. Participants performed abstractness ratings, imageability ratings, and CA ratings of all 36 words. The words were presented in random order in each task. All the ratings were made on a 7-point scale.

The predictions of the dual-coding theory were tested by having participants rate the imageability of each entity. For CA, we used instructions used in previous research. We asked participants to rate how difficult it would be to mentally generate a context for the entity. For all tasks, participants were encouraged to make ratings according to their personal understanding of the words.

Results

Manipulation Check

The word sample was constructed based on the MRC2 abstractness ratings. We checked whether the perception of our participants agreed with these abstractness ratings. Abstractness ratings performed by the participants were highly correlated with the MRC2 abstractness ratings (r = 0.94, p < 0.001). This indicates that participants in our study in fact perceived entities selected as most abstract as most abstract, and the most concrete entities as most concrete.

Predicting Abstractness of Overall Sample

Multiple regression analyses for the three predictors found that CA (r = 0.66) and IM (r = 0.77) both predicted concreteness ratings for the entire sample (p < 0.01), and for the *concrete* sub-sample. Ratings for these two variables were also significantly different for abstract versus concrete words in t-tests (t (34) = 4.41, p < 0.01 for CA and t (34) = 6.33, p < 0.01 for IM.) The number of contextual constraints was not a significant predictor; however, the percentage of abstract constraints was a good predictor (r = 0.47, p < 0.05).

Some of the predictor variables, especially ratings for CA and IM, were highly correlated. Therefore, a stepwise regression with all predictors was performed to examine their relative contribution towards explaining the variance in abstractness ratings. Only 23 cases were valid for this analysis because for some entities, none of the contextual constraints applied.

Predictors in this analysis included (i) CA ratings, (ii) IM ratings, (iii) the number of contextual constraints (CC), and the percentage of abstract constraints (ACC). The two highest predictors, which both contributed significantly to the regression, were IM and ACC. Together, these variables explained more than half of the variance (R^2 = 0.56). The change in the amount of variance explained by IM was 0.36 (F (1, 21) = 11.99, p < 0.01); the change due to ACC was 0.2 (F (1, 20) = 8.89, p < 0.01). The other variables did not add any significant changes in the amount of variance explained.

Predicting Abstractness of Abstract Sample

The only substantial predictor for ratings on the abstract nouns was the percentage of abstract contextual constraints (marginally significant; r = -0.47, p = 0.52). This measure was computed as the percentage of constraints for an entity that are not directly observable, such as mental / introspective constraints and relations.

This suggests that constraints play an important role for abstract entities over and beyond CA and IM. However, the number of words in this study was very limited. A second study was conducted to examine whether this finding holds up for a large set of abstract entities. Further, the second study involved entities of relatively high abstractness only, with less variance than in Study 1. Abstractness ranged from 2.2 to 3.6 on a 7-point scale, with a mean of 2.88 (SD=0.35). The result is a critical test of whether contextual constraints are discriminating enough to predict variation at such a fine level.

Study 2: Constraints and Abstractness

A total of 121 abstract nouns were coded for the contextual constraints described above. The coding indicated whether each constraint was *by necessity* part of a context in which the entity could occur. For example, *determination* requires an agent (who is determined), an agent goal, a certain attitude, and a stretch of time during which the attitude and goal do not vary (a state). Coding reliability on a 25% subset of the words, measured as correlation and Cohen's kappa, was significant (p<0.01) for three independent coders.

Based on the codes, we computed the number of constraints and the percentage of abstract constraints for each noun. Additionally, the codes were summarized across types of constraints to test whether particular kinds of constraints would yield particularly strong or weak predictions. The constraint groups were (1) concrete entities, (2) temporal constraints, (3) relational constraints, and (4) introspection-related constraints.

Results

All measures were submitted to a correlation with abstractness ratings from the MRC2 database. The astonishing result was that only the group of introspection-based constraints was significantly correlated (r=-0.21, p<0.05). This group includes mental constraints (goals, feelings, attitudes and beliefs, knowledge, and thoughts) and relational constraints between agents and other agents, objects, or issues. The percentage of abstract constraints yielded the second highest correlation (r=-0.14), but it was not significant. The finding for introspective constraints is interesting because it supports the recent proposal by Barsalou (1999) that introspection plays a central role in the processing of abstract concepts.

Overall, the result is consistent with the finding in study 1 that constraints play a role in our perception of abstractness. Importantly, the second study shows that constraints -- at least some of them – are good predictors even when fine discrimination is required.

Discussion

The results show that IM and CA are limited in explaining abstractness variations. Whereas ratings for both can account for the differences between abstract and concrete in general, they do not explain the variance in abstractness for different abstract concepts. This replicated the finding by Altarriba et al. (1999). The number of contextual constraints did not significantly predict abstractness ratings, but the abstractness of the constraints was a relatively strong predictor.

The findings are consistent with the two-factor model of abstractness, according to which abstractness and concreteness are determined by two different kinds of information. For concrete entities, both CA and IM were good predictors of the variation of ratings. For abstract concepts, the most critical type of information was the type of contextual constraint involved.

Research focusing on abstract concepts, rather than mere differences between concrete and abstract concepts, should presumably not be conducted on the basis of the assumptions made by CA and IM alone, since apparently these theories account mostly for the differences of the dichotomous classes of abstract and concrete entities. Research aiming to reveal characteristics of abstract entities should additionally take into account information that is specifically relevant to them, such as a system of context constraints that may guide our identification of these entities in context. The present results suggest that introspective processes and information, and to some extent abstract constraints in general, may be a good candidate for this abstract concept-specific information.

A system of constraints such as the one presented in Table 1 may further be useful in investigations of context effects, as they have been reported in previous research on abstract concepts. For example, some studies on CA effects produced the inconsistent result that abstract concepts were not processed faster when presented in context. A principled way to predict *what* context information is relevant to an abstract entity may account for such findings: Perhaps, the abstract words were presented in contexts that did not instantiate the relevant constraints.

Concrete versus Abstract

Maybe one of the most interesting implications of the findings is that constraints put the relevant information *outside of* the rated entity. That is, it is not an aspect of the entity itself that makes it abstract, but it is the abstractness of the constraints on situations in which it is used. This point offers a nice bridge to the CA theory, which argues that abstract concepts are abstract

because less context for their processing is available in memory.

The constraints may offer an explanation for this phenomenon. The more abstract the constraints are, the less guidance we have in constructing a mental context (or a simulation, see Barsalou, 1999). The constraints are there but they leave open most aspects of the concrete context. For example, the concept *comparison* requires (among other abstract constraints) the presence of two entities to be compared. The constraint does not dictate these entities to be of any particular nature, thus, they could be people, essays, houses, laws, feelings, etc.

In the case of a less abstract entity, such as *arrival*, the constraints involved are of a more concrete nature, and thus more effectively constrain the number of contexts we could construct to process the concept. An arrival involves an agent, an action (movement), and a particular location that the agent moves towards. These constraints can readily be used to simulate a fairly concrete situation in which an arrival takes place.

Features and Constraints

An interesting thought to pursue in future research is that features may fulfill the same function for concrete entities as abstract contextual constraints do for abstract entities. Thus, it is possible that the number of features / concreteness of features decrease from the concrete to the abstract pole, whereas the abstractness of the contextual constraints decreases from the abstract to the concrete end.

Familiarity Effects

This study controlled for familiarity effects through strategic sampling. Future research could examine an interesting question in connection to CA: It is likely that highly familiar concepts are represented with default contexts, which can easily be accessed, whereas contexts for other concepts have to be constructed.

Acknowledgments

Many thanks to Xu Xu, Jan Krug, and Priscilla Fernandez for assistance with material construction, data collection, and constraint coding, and to Larry Barsalou for productive discussions of some of the ideas reported here.

References

Altarriba, J., Bauer, L.M., & Benvenuto, C. (1999). Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behavior Research Methods, Instruments, & Computers, 31,* 578-602.

Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22,* 577-609.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology, 33A,* 497-505.

Crystal, D. (1995). The Cambridge encyclopedia of the English language. Cambridge: Cambridge University Press.

Gilhooly, K.J., & Logie, R.H. (1980). Age of acquisition, imagery, concreteness, familiarity and ambiguity measures for 1944 words. *Behaviour Research Methods and Instrumentation, 12,* 395-427.

Kacinik, N., Shears, C., & Chiarello, C. (2000). Familiarity for nouns and verbs: not the same as, and better than, frequency. In L.R. Gleitman & A.K. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (p. 1035). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kucera and Francis, W.N. (1967). Computational Analysis of Present-Day American English. Providence: Brown University Press.

Nelson, D.L., & Schreiber, T.A. (1992). Word concreteness and word structure as independent determinants of recall. *Journal of Memory and Language, 31,* 237-260.

Paivio, A. (1986). *Mental representations: A dual coding approach.* Oxford, UK: Oxford University Press.

Paivio, A., Yuille, J.C., & Madigan, S.A. (1968). Concreteness, imagery and meaningfulness values for 925 words. *Journal of Experimental Psychology Monograph Supplement, 76* (3, part 2).

Schank, R.C., & Abelson, R. (1977). *Scripts, plans, goals, and understanding*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schwanenflugel, P. (1991) Contextual constraint and lexical processing. In G. B. Simpson (Ed.), *Understanding word and sentence*. Amsterdam: Elsevier.

Schwanenflugel, P.J., & Shoben, E.J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 82-102.

Toglia, M.P. and Battig, W.R. (1978). Handbook of Semantic Word Norms. New York: Erlbaum.

Wiemer-Hastings, K., & Graesser, A.C. (1998). Abstract noun classification: A neural network approach. *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 1036-1042). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rules for Syntax, Vectors for Semantics

Peter Wiemer-Hastings (Peter.Wiemer-Hastings@ed.ac.uk)
Iraide Zipitria (iraidez@cogsci.ed.ac.uk)
University of Edinburgh
Division of Informatics
2 Buccleuch Place
Edinburgh EH8 9LW Scotland

Abstract

Latent Semantic Analysis (LSA) has been shown to perform many linguistic tasks as well as humans do, and has been put forward as a model of human linguistic competence. But LSA pays no attention to word order, much less sentence structure. Researchers in Natural Language Processing have made significant progress in quickly and accurately deriving the syntactic structure of texts. But there is little agreement on how best to represent meaning, and the representations are brittle and difficult to build. This paper evaluates a model of language understanding that combines information from rule-based syntactic processing with a vector-based semantic representation which is learned from a corpus. The model is evaluated as a cognitive model, and as a potential technique for natural language understanding.

Motivations

Latent Semantic Analysis (LSA) was originally developed for the task of information retrieval, selecting a text which matches a query from a large database (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990)[1]. More recently, LSA has been evaluated by psychologists as a model for human lexical acquisition (Landauer & Dumais, 1997). It has been applied to other textual tasks and found to generally perform at levels matching human performance. All this despite the fact that LSA pays no attention to word order, let alone syntax. This led Landauer to claim that syntax apparently has no contribution to the meaning of a sentence, and may only serve as a working memory crutch for sentence processing, or in a stylistic role (Landauer, Laham, Rehder, & Schreiner, 1997).

The tasks that LSA has been shown to perform well on can be separated into two groups: those that deal with single words and those that deal with longer texts. For example, on the synonym selection part of the TOEFL (Test of English as a Foreign Language), LSA was as accurate at choosing the correct synonym (out of 4 choices) as were successful foreign applicants to US universities (Landauer et al., 1997). For longer texts, Rehder et al (1998) showed that for evaluating author knowledge, LSA does steadily worse for texts shorter than 200 words. More specifically, for 200-word essay segments, LSA accounted for 60% of the variance in human scores. For 60-word essay segments, LSA scores accounted for only 10% of the variance.

In work on judging the quality of single-sentence student answers in an intelligent tutoring context, we have shown in previous work that although LSA nears the performance of intermediate-knowledge human raters, it lags far behind expert performance (Wiemer-Hastings, Wiemer-Hastings, & Graesser, 1999b). Furthermore, when we compared LSA to a keyword-based approach, LSA performed only marginally better (Wiemer-Hastings, Wiemer-Hastings, & Graesser, 1999a). This accords with unpublished results on short answer sentences from Walter Kintsch, personal communication, January 1999.

In the field of Natural Language Processing, the eras of excessive optimism and ensuing disappointment have been followed by study increases in the systems' ability to process the syntactic structure of texts with rule-based mechanisms. The biggest recent developments have been due to the augmentation of the rules with corpus-derived probabilities for when they should be applied (Charniak, 1997; Collins, 1996, 1998, for example).

Unfortunately, progress in the area of computing the semantic content of texts has not been so successful. Two basic variants of semantic theories have been developed. One is based on some form of logic. The other is represented by connections within semantic networks. In fact, the latter can be simply converted into a logic-based representation.

Such theories are brittle in two ways. First, they require every concept and every connection between concepts to be defined by a human knowledge engineer. Multi-purpose representations are not feasible because of the many technical senses of words in every different domain. Second, such representations can not naturally make the graded judgements that humans do. Humans can compare any two things (even apples and oranges!), but aside from counting feature overlap, logic-based representations have difficulty with relationships other than subsumption and "has-as-part".

Due to these various motivations, we are pursuing a two-pronged research project. First, we want to evaluate the combination of a syntactic processing mechanism with an LSA-based semantic representation as a cognitive model of human sentence similarity judgements. Second, we are

[1] We do not describe the functioning of the LSA mechanism here. For a complete description, see (Deerwester et al., 1990; Landauer & Dumais, 1997)

implementing a computational system to automate the processing of texts with this technique. This paper describes the human data we collected for the cognitive modeling aspect, the evaluation of our approach with respect to that data, and our initial attempts to implement the computational system.

Data collection

In (Wiemer-Hastings, 2000), we reported our initial attempts in this direction. In that evaluation, we compared our technique (described more fully below) to human ratings that were previously collected as part of the AutoTutor project (Wiemer-Hastings, Graesser, Harter, & the Tutoring Research Group, 1998). To our surprise, we found that adding syntactic information actually hurt the performance of an LSA-based approach. This could have been due to some problem with the approach, or due to some difficulty with the human data. The previous ratings had been based on complete multi-sentence student answers and ideal good answers. The raters were asked to indicate what percentage of the content of the student answer matched some part of the ideal answer. In the current work, we are looking at similarity ratings for specific sentences. Thus the previous data was not appropriate for our current goals.

To get more relevant human data, we started with text from the AutoTutor Computer Literacy tutoring domain so that we could more directly compare the results with our previous results, and because we had already trained an LSA space for it. AutoTutor "understands" student answers by comparing them to a set of target good answers with LSA. For this evaluation, we randomly paired 300 student answer sentences with 300 target good answers from the relevant questions. We split these into four booklets of 75 pairs, and gave each booklet to four human raters. Because we are also interested in the effect of knowledge on the reliability of ratings, we had previously asked the raters if they were proficient or not with computers. We gave each booklet to two proficient and two non-proficient raters.

We told the raters that the sentence pairs were from the domain of computer literacy, and asked them to indicate how similar the items were on a 6-point scale, from completely dissimilar to completely similar. Here is an example item:

> Sentence 1: ROM only reads information from the disk.
> Sentence 2: The central processing unit, CPU, can read from RAM.

We did not specify how the raters were to decide what similarity means.

The averaged correlations between the human raters were:

> Non-Proficient: $r = 0.35$, $P < 0.001$
> Skilled: $r = 0.45$, $P < 0.001$
> Mean Non-Proficient with Mean Proficient: $r = 0.53$, $P < 0.001$

Although these numbers are relatively low for inter-rater reliability on similarity tasks in general (Tversky, 1977; Goldstone, Medin, & Halberstadt, 1997; Resnik & Diab, 2000, for example), we have found this level of agreement in our other studies of sentence similarity. This task is obviously a difficult one for humans. In future work, we will study the effects of varying the level of context that is available for making these decisions.

Experiment 1: Part-of-speech tags

One way of deriving structural knowledge from text is by performing part of speech tagging. This is one area in which NLP systems have been fairly successful. Brill's tagger (Brill, 1994) is trained on a marked corpus and uses rules to assign a unique tag to each word. It first assigns the most common tag for each word, then applies a set of automatically-derived context-based rules to modify the original tagging.

When LSA is trained, it does not distinguish between words which are used in multiple parts of speech. This may have significant semantic ramifications. The word "plane", for example, has very different senses as a verb and as a noun. One way to add structural information to LSA would be to allow it to distinguish the part of speech for each word when training and comparing sentences.

Approach

Our approach to this task was to use the Brill tagger to assign a part-of-speech tag to every word in the training corpus and every word in the test set (which had been given to the human raters). The tag for each word was attached to it with an underscore so that LSA would view each word/tag combination as a single term. For example:

```
ROM_NNP is_VBZ information_NN the_DT
computer_NN was_VBD programmed_VBN with_IN
when_WRB it_PRP was_VBD built_VBN ._.
```

The original training corpus was 2.3 MB of text taken from textbooks and articles on computer literacy. We trained LSA on the tagged corpus at 100, 200, 300, and 400 dimensions because these dimensions had shown reasonable levels of performance in previous evaluations. Then we evaluated this approach by using the new LSA space to calculate the cosines between the tagged versions of the test sentences that had been given to the human raters. We calculated the correlations between the cosines and the human ratings.

Results

Figure 1 shows the correlations between the different LSA models and the human ratings. The first bar depicts the correlation using the standard LSA space (at 200 dimensions) as applied to the untagged versions of the sentences.

Discussion

It is clear that the performance of the tagged models do not match human judgements as closely as the standard

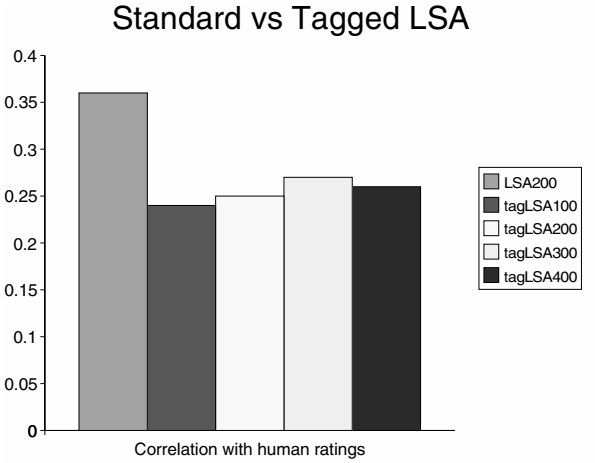


Figure 1: The performance of tagged LSA

LSA approach does. It is not clear why this is. This could support Landauer's claim with respect to the irrelevance of structural information for determining meaning. Perhaps LSA somehow does manage to account for different senses and uses of a word even though it does not have explicit knowledge of the syntactic context of the word's use.

Alternatively, the relatively poor performance could be due to some inadequacies of this particular approach. For example, although the number of dimensions was in the correct relative range with respect to non-tagged LSA processing, perhaps tagged LSA works better with more (or fewer?) dimensions. It could also be that the performance was hampered by mistagging of key words in the sentence. Because the Brill tagger is trained on the Wall Street Journal corpus, its tagging rules often lead it astray when processing the colloquial and domain-specific student answers in our tutoring domain. For example, one student answered a question about a computer's memory like this, "RAM stores things being worked with." The Brill tagger mistagged the word "stores" as a plural noun, thus greatly altering the overall meaning of the sentence.

Experiment 2: Surface parsing

Another obvious potential contribution of sentence structure to meaning is by providing information about the relationships and actions of the participants: the "who did what to whom" information. Although some might claim that LSA is able to derive this information from its training corpus (because men rarely bite dogs, for example), this can not always be the case. And with the exception of case-marked pronouns like "I" and constructions like "there is . . . ", it is difficult to think of any entity references that can not appear as both subject and object of a sentence. Thus, if we can separately determine the subject, object, and verb parts of a sentence, we should be able to provide information that, in addition to that which we get from LSA, will improve sentence similarity judgements.

The approach: Structured LSA

In standard LSA, the input to the procedure is an entire text, represented as a string. The string is then tokenized into words, and the vector for each word is accessed from the trained vector space. LSA ignores words that it can not find, i.e. those that did not appear in more than one document in the training corpus, or those that appear on the stop-word list, a list of 440 very common words, including most function words. The vector for a text is constructed by simply adding together the relevant word vectors. Two texts are compared by calculating the cosine between their vectors.

In our approach which we call Structured LSA (SLSA), we preprocess input sentences to derive aspects of their structure. More specifically, for each sentence, we:

- resolve pronominal anaphora, replacing pronouns with their antecedents,
- break down complex sentences into simple sentences,
- segment the simple sentences into subject, verb, and object substrings.[2]

.

For example, we transform the student answer: "RAM stores things being worked with, and it is volatile" into:

("stores" "RAM" "things being worked with")
("volatile" "RAM")

This yields a verb string, subject string, and (optional) object string for each sentence. Note that for copular sentences as in the second simple sentence above, the "verb string" is the description following the verb. Also note that our human data was collected not on the original sentences, but on sentences on which the first two steps above were already completed.

SLSA similarity rating

To calculate a similarity score between two sentences with the SLSA approach, the preprocessing is performed on the sentences. Then we separately pass the verb strings, subject strings, and object strings to LSA which computes the cosines between them. Then we average the three together to get an overall similarity rating between the sentences.[3]

Note that this approach provides more information than the standard LSA approach. For each pair of sentences, there are four separate similarity ratings instead of just one.

[2]Passive sentences were normalized, putting the syntactic object as the subject, and vice versa.

[3]In (Wiemer-Hastings, 2000), we evaluated three different methods for aggregating the segment cosines, including a subject-predicate approach and given-new approach. In the current evaluation, the simple average provided the best performance, so we do not present the others here.

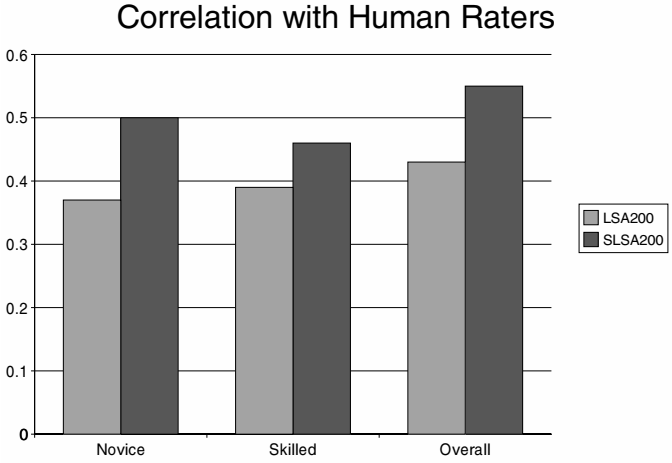


Figure 2: The correlation between SLSA scores and human ratings

In addition to the overall similarity score, SLSA produces separate measures of the similarity of the segments of the sentences. This additional information could be very useful for dialog processing systems.

Results

Because we were interested in evaluating this approach in principle, and not with respect to any particular implementation of the preprocessing technique, we preprocessed the entire test set by hand as described above. Then, the SLSA similarity scores were calculated and correlated with the human ratings. Figure 2 compares the LSA and SLSA approaches with respect to the correlations to human ratings for the non-proficient, proficient, and averaged ratings.

SLSA performed better with respect to each subset of human ratings than did the standard LSA approach. The correlation with the mean of all four human raters was slightly better than the highest level of agreement among the human raters.

Discussion

These results are not consistent with Landauer's claim that syntax does not convey additional semantic content beyond the meanings of individual words. Human sentence similarity judgements are better modelled by an approach that takes structural information into acccount. Although the standard LSA approach does perform as well as humans on longer texts, this may be because the information about who does what to whom in individual sentences is lost in the noise, or is constrained by the larger context.

Toward a hybrid natural language understander

Now that we have validated the benefits of this approach, we have begun to develop a system that will use shallow parsing techniques to automatically perform the preprocessing of input sentences for SLSA. This section describes different approaches that we are evaluating.

Surface parsing

Shallow parsing is currently an area of intense interest in the corpus-based natural language processing community. In fact, the 2001 Computational Natural Language Learning workshop at the Association for Computational Linguistics conference will include a shared task which is to evaluate different techniques for clause splitting. Clause splitting is defined as separating a sentence into subject and predicate parts.

We are currently evaluating the feasibility of using several publicly available surface parsing tools: LTChunk, the SCOL parser, and the Memory-Based Shallow Parser (MBSP). LTChunk was developed by the Language Technology Group at the University of Edinburgh (described at `http://www.ltg.ed.ac.uk/software/chunk/`). It identifies noun phrases and verb groups (combinations of adverbs, auxiliaries, and verbs) in text. The SCOL parser was developed by Abney (1996), and parses text using a set of cascaded rules which delay "difficult" decisions like where to attach prepositional phrases. MBSP (Daelemans, Buchholz, & Veenstra, 1999) is part of the Tilburg Memory Based Learner project (Daelemans, Zavrel, van der Sloot, & van den Bosch, 2000). It is also trained on Penn Treebank Wall Street Journal corpus, performs part-of-speech tagging, and segments texts into subject, verbs, and objects.

Current work

Each of these different methods has drawbacks. The corpus trained approaches have the same difficulty as that noted above: the student answer texts differ sufficiently from the Wall Street Journal to lead to many mistaggings, and therefore, misparses.

Our current efforts are focussing on using the Brill Tagger (adjusting its tags to be more appropriate for our domain), and then the SCOL parser to identify sentence segments. We are developing a postprocessor to transform the output of the parser into the subject, verb, object segmentation that we need as input to SLSA. The postprocessor handles active, passive, and imperative constructions. We are also working on a simple coreference resolution mechanism to allow substitution of antecedents. Our set of hand-processed sentences gives us a useful gold standard against which to evaluate our approach.

The process of matching the segments of the two sentences can be viewed as structure mapping of the type that Gentner et al developed for processing analogies (Gentner, 1983; Forbus, Ferguson, & Gentner, 1994, for example). Ramscar and colleagues have developed a two-stage model for processing analogy which first performs structure-mapping between two scenarios, and then uses LSA to determine the similarity of the slot fillers between the two structures (Yarlett & Ramscar, 2000). For SLSA, the proper treatment of syntactic structures like passives is

quite important. Even more difficult are alternations like "give" and "take" which can have the same syntactic structure, but completely different semantic role structures. Resolving such cases seems to require semantic information, resulting in a chicken-and-egg situation. How can we use SLSA to interpret the meaning of a sentence if we must know the meaning in order to use SLSA? Our current research involves treating the verbal and nominal parts of the input sentences differently.

Conclusions

Our findings do not support the claim that syntax provides a negligible contribution to sentence meaning. Instead, a sentence comparison metric that combines structure-derived information with vector-based semantics models human similarity judgements better than LSA alone. As previously mentioned, this approach provides a number of advantages. Its overall fit to human data is not only better than standard LSA, but it provides additional information about the similarity of the different parts of sentences. This could be used in a dialogue-based tutoring system to focus the student's attention on some particular aspect of the target good answer.

With respect to traditional parsing techniques, SLSA has three obvious advantages. First, it is fast, because it does not deal with the combinatorial explosions from ambiguity that most parsers face. Second, it does not require a hand-built semantic concept representation which is tedious to build and brittle. Third, LSA is (in a sense) grounded. Although it does not have direct experience of the world, LSA does have indirect experience via its training corpus. The corpus provides a rich set of interconnections between terms which allows LSA to successfully model many aspects of human linguistic competence.

The limitation of SLSA as a natural language understanding mechanism is that it is only appropriate for tasks where understanding can be cast as computing the similarity of an item to an expected utterance. For tutoring, the approach is feasible because the tutor (whether computer or human) normally determines the topic of conversation, and has some idea of what the student should say. For other tasks where the input utterance is less constrained, this approach might not be the best. On the other hand, if a natural language generation system could be used to generate a set of expected utterances in a particular domain, expectation-based understanding might be feasible and effective.

Although we have presented the syntactic analysis of this work as being derived from symbolic, rule-based mechanisms, our analyses of SLSA as a cognitive model do not depend on this. They would be equally applicable with a connectionist surface parsing technique.

These findings raise quite a few interesting questions for future research. For example:

- What exactly are humans measuring when they rate sentence similarity? Perhaps varying the instructions for human raters will get them to focus on different aspects of meaning.
- What is the best level of granularity to use in segmenting sentences? We have evaluated the use of subject, verb, and object segments, but a coarser or finer segmentation may perform better.
- How much semantic information can be derived from a sentence without knowing its meaning? Inducing additional relationships between the parts of a sentence might improve the SLSA approach, but may require already knowing what the sentence is about.

Addressing these questions will be the focus of our future research.

References

Abney, S. (1996). Partial parsing via finite-state cascades. In *Proceedings of the ESSLLI '96 Robust Parsing Workshop*.

Brill, E. (1994). Some advances in rule-based part of speech tagging. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*. AAAI Press.

Charniak, E. (1997). Statistical Parsing with a Context-free Grammar and Word Statistics. In *Proceedings of the 14th National Conference of the American Association for Artificial Intelligence, Providence, RI., July*, pp. 598–603.

Collins, M. (1996). A New Statistical Parser Based on Bigram Lexical Dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, Santa Cruz, CA*, pp. 184–191 San Francisco, CA. Morgan Kaufmann.

Collins, M. (1998). *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Daelemans, W., Buchholz, S., & Veenstra, J. (1999). Memory-Based Shallow Parsing. In *Proceedings of CoNLL-99*.

Daelemans, W., Zavrel, J., van der Sloot, K., & van den Bosch, A. (2000). TiMBL: Tilburg Memory Based Learner, version 3.0, Reference Guide. Tech. rep. Technical Report 00-01, 2000, ILK, University of Tilburg. available at http://ilk.kub.nl/.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, *41*, 391–407.

Forbus, K., Ferguson, R., & Gentner, D. (1994). Incremental structure mapping. In *Proceedings of the 16th Annual Conference of the Cognitive Science Society* Mahwah, NJ. Erlbaum.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155–170.

Goldstone, R., Medin, D., & Halberstadt, J. (1997). Similarity in context. *Memory and Cognition*, *25*(2), 237–255.

Landauer, T. K., Laham, D., Rehder, R., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, pp. 412–417 Mahwah, NJ. Erlbaum.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Rehder, B., Schreiner, M., Laham, D., Wolfe, M., Landauer, T., & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes*, *25*, 337–354.

Resnik, P., & Diab, M. (2000). Measuring Verb Similarity. In *Proceedings of the* 22^{nd} *Annual Conference of the Cognitive Science Society* Mahwah, NJ. Erlbaum.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

Wiemer-Hastings, P. (2000). Adding syntactic information to LSA. In *Proceedings of the* 22^{nd} *Annual Conference of the Cognitive Science Society*, pp. 989–993 Mahwah, NJ. Erlbaum.

Wiemer-Hastings, P., Graesser, A., Harter, D., & the Tutoring Research Group (1998). The foundations and architecture of AutoTutor. In Goettl, B., Halff, H., Redfield, C., & Shute, V. (Eds.), *Intelligent Tutoring Systems, Proceedings of the 4th International Conference*, pp. 334–343 Berlin. Springer.

Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. (1999a). How Latent is Latent Semantic Analysis?. In *Proceedings of the Sixteenth International Joint Congress on Artificial Intelligence*, pp. 932–937 San Francisco. Morgan Kaufmann.

Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. (1999b). Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In Lajoie, S., & Vivet, M. (Eds.), *Artificial Intelligence in Education*, pp. 535–542 Amsterdam. IOS Press.

Yarlett, D., & Ramscar, M. (2000). Structure-Mapping Theory and Lexico-Semantic Information. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pp. 571–576 Mahwah, NJ. Erlbaum.

Did Language Give Us Numbers?
Symbolic Thinking and the Emergence of Systematic Numerical Cognition

Heike Wiese (heike.wiese@rz.hu-berlin.de)
Humboldt-University Berlin, Department of German Language and Linguistics, Unter den Linden 6
10099 Berlin, Germany

Abstract

What role does language play in the development of numerical cognition? In the present paper I argue that the evolution of symbolic thinking (as a basis for language) laid the grounds for the emergence of a systematic concept of number. This concept is grounded in the notion of an infinite sequence and encompasses number assignments that can focus on cardinal aspects ('three pencils'), ordinal aspects ('the third runner'), and even nominal aspects ('bus #3'). I show that these number assignments are based on a specific association of relational structures, and that it is the human language faculty that provides a cognitive paradigm for such an association, suggesting that language played a pivotal role in the evolution of systematic numerical cognition.

Introduction

Over the last decades, results from several disciplines relating to cognitive science (in particular from psycholinguistics, developmental psychology, cognitive ethology, and cognitive neuroscience) have shed new light on the relationship between language and numerical cognition.

On the one hand, the acquisition of some aspects of mathematical knowledge seems to be linked to the number words of a language. Psychological and neurological studies suggest that the representation of memorised mathematical knowledge such as multiplication tables and its application in mental calculation is closely linked to the language it was originally learned in (cf. Dehaene, 1997).

In addition, cross-linguistic studies on the acquisition of number words have shown that the structure of a number word sequence can have an impact on children's mathematical performance:[1] a highly regular and transparent number word sequence makes it easier for children to grasp multiplicative and additive relationships between numbers and to correlate them with Arabic numerals, than a sequence that contains opaque elements.

For instance in the Chinese number word sequence, as opposed to the one in English, the underlying decimal structure is always transparent in complex number words (for instance, the Chinese counterparts for English 'ten – eleven – twelve – thirteen – fourteen – ... – twenty' have the form 'ten – ten-one – ten-two – ten-three – ten-four – ... – two-ten'). In accordance with this linguistic difference, Chinese children were shown to have a better grasp of the base ten structure of their number system and performed initially better in arithmetic tasks than their American counterparts.

On the other hand, converging evidence from developmental psychology and cognitive ethology has revealed numerical capacities that seem to be independent of language. Preverbal infants as well as higher animals were shown to be able to grasp small numerosities (the cardinality of small sets) and perform simple arithmetic operations on them.[2] Evidence from lesion and brain-imaging studies indicates that a specific brain region, the inferior parietal cortex, might be associated with this ability.[3]

This suggests that, while some later aspects of mathematical cognition might be influenced by linguistic factors, we also possess a biologically determined concept of cardinality: a concept of numerical quantities and their inter-relations that is independent of the acquisition of a specific language, and independent of the human language faculty in general.

Does this mean that our concept of *number* is independent of language? In this paper, I will argue that it is not. I will argue that language contributed to numerical cognition in a fundamental way: in the history of our species the emergence of language as a mental faculty opened the way for systematic numerical cognition. Symbolic thinking as the basis of language provided a cognitive pattern that enabled humans to make the step from primitive quantitative reasoning to a generalised concept of number, a concept that is not restricted to cardinality, but allows us to employ numbers to identify cardinal as well as ordinal and even nominal relationships between empirical objects.

To develop this claim, I will first spell out the relationship between numbers and cardinality and show that it is crucial for our understanding of the cognitive

[1] Cf., for instance, Miura et al. (1993) and Ho & Fuson (1998) for Asian (Chinese, Korean and Japanese) versus US-American (English-speaking) and European (British, French and Swedish) first-graders and kindergarteners.

[2] Cf. Wynn (1998) for a detailed discussion of the evidence from infants and new-borns; Butterworth (1999) and Dehaene (1997) for overviews of numerosity concepts in human infants and animals.

[3] Cf. Dehaene, Dehaene-Lambertz & Cohen (1998).

number domain not to focus on cardinality alone. I will then introduce a unified notion of number assignments that brings together cardinal, ordinal and nominal aspects. On this basis, I analyse structural parallels between number assignments and symbolic reference that suggest that language provides a cognitive pattern for systematic number assignments.

Numbers and Cardinality

One of the aspects that make numbers so interesting is their enormous flexibility. A quality like colour, for instance, can only be conceived for visual objects, so that we have the notion of a red flower, but not the notion of a red thought. In contrast to that, there seem to be no restrictions on the objects numbers can apply to. In his 'Essay Concerning Human Understanding', John Locke put it this way: "[...] number applies itself to men, angels, actions, thoughts; everything that either doth exist, or can be imagined." (Locke 1690, Book II, Ch.XVI, §1).

This refers to contexts where numbers identify the cardinality of a set: they tell us how many men or actions etc. there are in the set. This number assignment works for any sets of objects, imagined or existent, no matter what qualities they might have otherwise; the only criterion is here that the objects must be distinct in order to be quantified.[4]

Frege (1884) regarded this flexibility as an indication for the intimate relationship between numbers and thought: "The truths of arithmetic govern all that is numerable. This is the widest domain of all; for to it belongs not only the existent, not only the intuitable, but everything thinkable. Should not the laws of number, then, be connected very intimately with the laws of thought?" (Frege 1884, §14).

However, this is only one respect in which numbers are flexible. Not only can we assign them to objects of all kinds, we can also assign them to objects in ways that are so diverse that on first sight, they seem not to be related at all. Of these number assignments, the one that relates to the cardinality of sets is probably the first that comes to mind, but it is by no means the only way we can assign numbers to objects.

The same number, say 3, can be used to give the cardinality of pencils on my desk ('three pencils'); to indicate, together with a measure unit, the amount of wine needed for a dinner with friends ('three litres of wine'); it can tell us the rank of a runner in a Marathon race ('the third runner'); or identify the bus that goes to the opera ('bus #3' / 'the #3 bus').[5]

[4] This criterion on objects can be reflected in language by the distinction of count nouns versus mass nouns, as their designations (cf. also Wiese & Piñango, this volume).

[5] As the examples in brackets illustrate, these different usages of numbers establish different contexts for number words that have to be mastered in first language acquisition. Cf. Fuson & Hall (1983) for a study of the acquisition process.

We can subsume our different usages of numbers under three kinds of number assignments: cardinal, ordinal, and nominal assignments (cf. Wiese, 1997).

Cardinal number assignments are denoted by expressions like 'three pencils' or 'three litres of wine', where 'three' is an answer to 'How many?'. In cardinal assignments, the number identifies the cardinality of a set, e.g. a set of pencils or a set of measure units that identify a certain volume (in our example, litres).

In *ordinal number assignments*, the number applies to an element of a sequence. For instance in the Marathon example, 3 indicates the rank of a particular person within the sequence of runners (the third runner).

We encounter *nominal number assignments* in the form of house numbers, in subway and bus systems, in the numbering of football players, or in telephone numbers. What these cases have in common is the fact that the numbers identify objects within a set: in nominal assignments, numbers are used as readily available (and inexhaustable) proper names. So rather than thinking of names like 'Mike' or 'Lucy' for buses, we assign them numbers when we want to identify them (for instance, 'bus #3'), and similarly, we assign numbers to houses in a street or to the members of a football team.

Hence, numbers are flexible tools that can be used in a wide variety of contexts, where they identify different properties of objects. Of these properties, cardinality is only one instance – it is *a* property that we can identify with numbers, but it is not necessarily more closely connected to numbers than other properties that can also be identified in number assignments (that is, the rank of an object in a sequence, or the identity of an object within a set). Figure 1 illustrates this view:

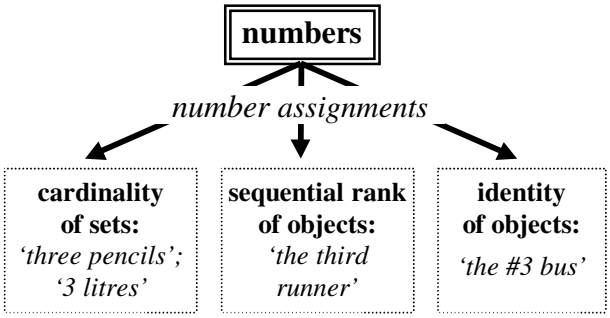


Figure 1: Numbers as flexible tools –
Integration of cardinality into the number domain

This approach, then, integrates cardinality into a broader view of the number domain. It distinguishes numbers and cardinality by characterising numbers as elementary tools that are not necessarily linked up with cardinality, but can equally bear on cardinal aspects, ordinal aspects, or nominal aspects in application, when employed in the different kinds of number assignments.

A Unified Approach to Number Assignments

What is it that makes numbers so flexible, how are their different usages related to each other? To answer this question, let us have a closer look at the different ways numbers apply to objects, that is cardinal, ordinal and nominal number assignments.

A theory that gives us a handle on these different types of number assignments is the Representational Theory of Measurement (henceforth, RTM).[6] This theory, which has been highly influential within philosophy and experimental psychology, is concerned with the features that make a number assignment[7] significant; it aims to establish the criteria that make sure that the number we assign to an object does in fact tell us something about the property we want to identify.

In the present section I will employ the machinery of this theory to a somewhat different purpose, interpreting the RTM as a unified framework for number assignments. This framework allows us to lay down the constitutive features of meaningful number assignments, the features that underly a systematic concept of numbers and of the relations which they identify between empirical objects.

In a preliminary approach, we can identify an assignment of numbers to objects as meaningful when certain relations between the numbers represent relations between the objects. Figure 2 gives an example:

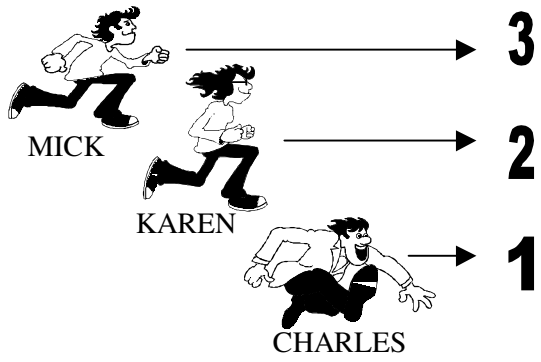


Figure 2: A meaningful number assignment: Numerical ranking of runners in a race

In this instance of number assignments, numbers have been assigned to participants in a race, such that the '<' relation between the numbers represents the ordering of the runners by the relation 'is faster than': Charles, as the fastest runner, received the smallest number, 1; Mick, who is slowest, received the largest number, 3, and Karen, who is faster than Mick and slower than Charles, got a number that is smaller than Mick's and larger than Charles's, namely 2. This way the ordering of the runners as 'Charles is faster than Karen who is faster than Mick' is reflected by the ordering of the numbers that they received: '1 < 2 < 3'.

The general features that make a number assignment meaningful can be captured by two requirements. The first requirement is that we regard the objects and the numbers only insofar as they form *relational structures*, that is, sets of elements that stand in specific relationships to each other. The two relational structures are distinguished as *numerical relational structure* (the relational structure constituted by the numbers) and *empirical relational structure* (the one established by the objects).

Accordingly, in the runner example we regarded the runners not as unrelated individuals, but treated them as elements of a particular sequence. The empirical relational structure is here constituted by the relation 'is faster than'. The relation between the numbers that we focused on was '<' ('lesser than'). All other relations that might hold between the objects (for example, the relative age of the runners) or between the numbers (for example, odd numbers versus even numbers), are ignored for the purposes of number assignment.

The second requirement for the number assignment is that the correlation between numbers and objects constitutes a *homomorphic* mapping, one that not only correlates the elements of the two relational structures, but also preserves the relevant relations between them.

In our example, the homomorphism associates the relation 'runs faster than' from the empirical relational structure (the sequence of runners) with the '<' relation in our numerical relational structure (the numbers). So for instance from the fact that one runner received the number 2 and another one got the number 3, one can deduce that the first runner was faster than the second one, because 2 < 3.

The interesting aspect for our discussion is now that this implies that number assignments are essentially links between relations: it is not so much the correlation between individual objects and individual numbers that counts, but the association of relations that hold between the empirical objects with relations that hold between the numbers.

As a result, ew can now analyse the different kinds of number assignments as instances of a unified pattern: they are constituted by a homomorphic mapping between two relational structures; a mapping that associates, in each case, a particular numerical relation with a relation between empirical objects:[8]

---

[6] Cf. Krantz et al. (1971), Narens (1985), Roberts (1979).

[7] The RTM uses the term 'measurement' (instead of 'number assignments') here. This terminology is slightly at odds with our pre-theoretical usage, where 'measurement' refers only to a particular class of cardinal number assignments (those identifying empirical properties like weight, volume or temperature), but excludes ordinal and nominal number assignments, which are included under the RTM notion of 'measurement'. In the present paper, I therefore use the more intuitive term '(meaningful) number assignments'.

[8] For a detailed discussion and formalisation of the different kinds of number assignments cf. Wiese (2001).

In *cardinal number assignments*, the empirical objects are sets. A number *n* identifies the cardinality of a set *s* (*n* tells us how many elements *s* has). The mapping associates the numerical relation '>' with the empirical relation 'has more elements than'. The number assignment is meaningful if and only if a one-to-one-correlation between the numbers from 1 to *n* and the elements of *s* is possible. For instance, when we assign the number 3 to a set of pencils, this number assignment can be regarded as a meaningful cardinal number assignment when it is possible to link up each pencil with a different number from 1 to 3. (We employ this verification procedure in counting routines.)

In *ordinal number assignments* (like the one in Figure 2), the empirical objects are not sets, but individual elements of a sequence. A number *n* identifies the rank of an object within a sequence *s*. For this task, we focus on the sequential order of numbers. The homomorphism that constitutes our number assignment associates the numerical relation '<' (or '>', respectively) with the relative ranks of the objects within *s* (for instance, the relative ranks of runners as established by the relation 'is faster than' in Figure 2). The number assignment is meaningful if and only if objects receive higher and lower numbers with respect to their higher and lower positions within *s*.

In *nominal number assignments*, the empirical objects are elements of a set (for example the bus lines in a city), and the numbers are used as labels: a number *n* identifies an object within a set *s*. The mapping associates the numerical relation '=' (or '≠') with the empirical relation 'is (non-)identical with'. The numerical statement is meaningful if and only if distinct objects always receive distinct numbers.

What these different number assignments have in common is the translation of relational structures. In cardinal, ordinal, and nominal number assignments alike, a relation between empirical objects is associated with a relation that holds between the numbers. It is this translation of relational structures that constitutes number assignments, and by doing so, lays the ground for systematic numerical cognition.

How did this principle evolve? In the following section I argue that the translation of relational structures as a cognitive pattern might have its origins in the emergence of symbolic thinking. I will argue that it is symbolic thinking, as a basis for the human language faculty, that made this pattern available to the human mind, and this way enabled us to develop a systematic number concept.

The Contribution of Language to the Emergence of a Systematic Number Concept

According to an account of language evolution as developed in Deacon (1997), the main step in the emergence of human language (as opposed to animal communication systems) is the development of a symbolic system; in a process of co-evolution of language and the brain, the adaptation of our brain to symbolic thinking gave rise to the emergence of the linguistic faculty we have today. To understand the significance of this view for our investigation into numbers and language, it is crucial to understand what Deacon means by *symbolic reference* here.

Following a semiotic taxonomy as introduced by Charles Peirce, Deacon distinguishes three kinds of signs: icons, indices and symbols. In *iconic* reference the sign shares some features with its referent, it is similar to the object it refers to (such as the icon ♿ that refers to a wheel-chair user). In *indexical* reference the sign is related to the object by a physical or temporal relation; it occurs together with its referent (for instance, tears could be interpreted as an index for grief).

In *symbolic* reference, the link between sign and object is established by convention, as in the case of human languages. The critical similarity, the similarity between symbols and their referents, emerges on a higher level, namely on that of the system. Symbols are always part of a system, and they refer to objects not as individual tokens, but with respect to their position in that system. In the case of symbols, reference shifts from individual signs and individual objects to relations between signs and relations between objects; it shifts from the token to the system.

Under this account, symbolic reference as the basis of human languages is crucially a link between relations (sign-sign and object-object), not between individuals (signs and objects). It is the relations between words that reference is based on.

These can be linear relations like the order of words in a sentence, or hierarchical relations like 'object of' or 'subject of', which mark the relations between a verb and its complements. For instance in the sentence "The dog bites the rat." one can identify the dog as the attacker and the rat as the victim, because the noun phrase 'the dog' comes before the verb, which is the position for the subject in English, and 'the rat' comes after the verb, in object position, and the noun phrases in these positions denote the Agent (attacker) and the Patient (victim) of the 'biting'-action, respectively.

So the connection one makes is between (a) symbolic relations like 'The words *the dog* come before the word *bites*' (linear) or 'The noun phrase *the dog* is subject of the verb *bite*' (hierarchical) and (b) relations between referents, namely 'The dog is the Agent in the biting-event'; and similarly for the rat:

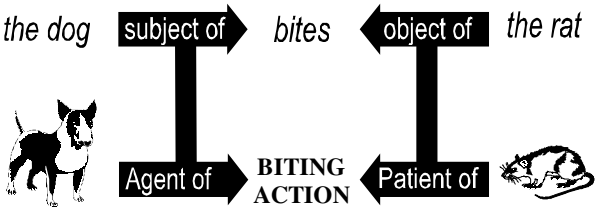


Figure 3: Symbolic reference as an association of relations

According to Jackendoff (1999), in the course of language evolution direct relationships between the linear order of words and their referents are replaced by links that are mediated by complex hierarchical syntax. Here, this would mean that the focus of symbolic relationships shifts from linear to hierarchical relations: on an early stage in the evolution of language, *linear relations* between symbols as evident in speech ('comes before / after') would directly be associated with hierarchical relations between referents ('Agent of / 'Patient of'), whereas on a later stage we would have (syntactic) *hierarchical relations* between symbols, like 'subject of / object of', which can be linked up with hierarchical relations between referents.

In both cases, it is the relationships that are associated, rather than individual symbols and individual referents. Unlike in iconic and indexical reference, in symbolic reference we pick out an object indirectly, relying on links that connect relationships between symbols (such as 'comes before / comes after', or 'subject of` / object of') with relationships between objects (such as 'Agent of / Patient of'). This is what symbolic reference is ultimately about: it is a connection between signs and referents that focuses on relationships.

This means that symbolic reference is constituted by a mapping between *relational structures*: we regard the symbols and their referents only insofar as they are part of a system whose elements stand in specific relations to each other; the association of symbols and their referents is determined by the respective relations that hold between them.

This is a phenomenon very similar to the one we encountered in the case of number assignments. As Figure 4 illustrates (for two of the runners from Figure 2 above), number assignments are based on links between relations, too: in number assignments we associate numerical relations with relations between empirical objects, just as in language we associate *symbolic* relations with relations between objects.

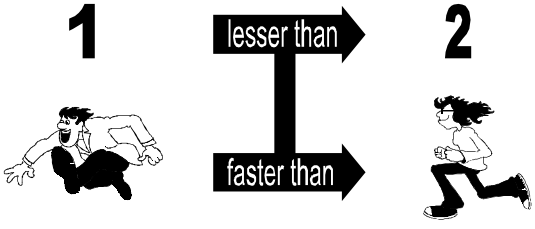


Figure 4: Number assignment as an association of relations

When we assign numbers to empirical objects, the links we establish are not between individual numbers and individual objects, but between a numerical relational structure and an empirical relational structure. And when we assign symbols to their referents, the links we establish are not between individual signs and individual objects, but between a relational structure of signs and a relational structure of the objects that they refer to.

This means that we can identify the same pattern in number assignments and in symbolic reference: in number assignments a numerical relational structure is correlated with an empirical relational structure; in symbolic reference a 'symbolic relational structure' is correlated with an empirical relational structure.

In both cases, the links between individual tokens (a number and an object, or a symbol and its referent) are based on their respective positions in the system, they are constituted by links between relations (numerical relations and empirical relations, or symbolic relations and relations between referents). Figure 5 illustrates these parallels:

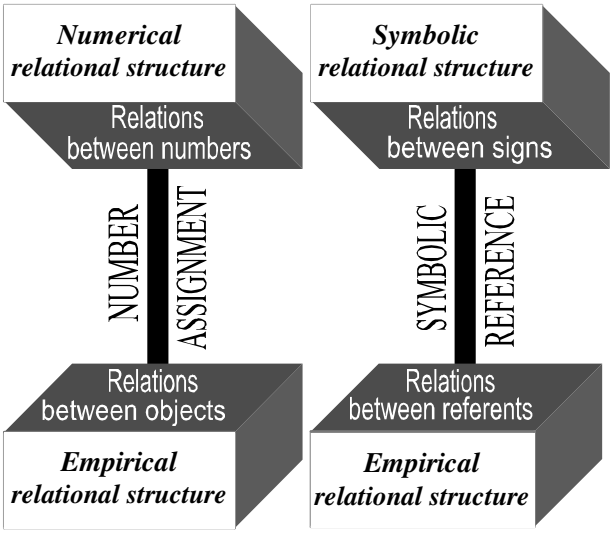


Figure 5: Translation of relational structures in number assignments and symbolic reference

This puts number assignments in a close association with the symbolic reference that lies at the core of our linguistic capacity, and shows us a way how systematic numerical cognition could have evolved in the human mind: in the development of our species the evolution of symbolic thinking in the emergence of language might have enabled us to grasp the logic of number assignments.

Once we passed the symbolic threshold, a paradigm was set for the systematic correlation of relational structures, and could be applied in the number assignments that underlie our numerical concepts. This way symbolic thinking prepared the way for systematic numerical cognition.

Under this approach, we can account for the capacity to systematically assign numbers to objects by a relatively small evolutionary step. According to this account, the use of numerical relational structures did not develop from scratch, but could build on already existing cognitive patterns that had evolved as part of sym-

bolic cognition – a re-usage that makes a lot of sense in terms of evolutionary economy.

At the same time, language gave us a handle on infinity. The phrases we can potentially generate in a language represent a discrete infinity: from a set of primitive elements – the lexical items of our language – we can generate an infinite number of complex constructions by means of combinatorial rules. In the words of Steven Pinker: "In a discrete combinatorial system like language, there can be an unlimited number of completely distinct combinations with an inifinite range of properties." (Pinker 1994, 84).

It is these combinatorial rules that constitute the infiniteness of number word sequences. The sequences of words we employ for counting ('one, two, three, …') are open-ended because of the generative rules governing the construction of complex elements. Through number words, language provides us with the notion of an infinite sequence.[9]

Note that it is the possesion of the *language faculty*, the emergence of language as a mental faculty in the history of our species, that is crucial here, not the successful and complete acquisition of a particular language in individual development. This also means that acquired or innate impairments of the language capacity do not necessarily affect our ability to grasp number assignments, as long as the basic linguistic capacity is still intact (including the association of relational structures by homomorphic mappings).

And let me emphasise again that this does not mean that without language, we would have no concept of properties like cardinality or rank that we identify with numbers. As the above-mentioned evidence from animal studies and studies with human infants shows, the emergence of our number concept could draw on prelinguistic capacities we share with other species, for instance our grasp of cardinality as a property of sets.

Language has been crucial in integrating these early concepts into a systematic number concept, one that is based on an infinite sequence of numerical tools that can be used to identify empirical properties via a correlation of relational structures.

Under this notion, numerical cognition as well as language can be regarded as genuinely human; as mental faculties that are not merely of greater complexity (than, say, animal communication systems and numerosity concepts) and grounded in a higher general-purpose intelligence, but qualitatively different and specific to the human mind.

---

[9] Cf. Hurford (1987) for a detailed analysis of number words; Wiese (1997; 2001) for the status of number word sequences within language and numerical cognition. In Wiese (2001, ch.4) I show that linguistic generativity (and therefore infinity) could be passed on to numerical cognition via counting sequences, and that this transfer could take place not only in individual cognitive development – as for instance assumed by Bloom (1994) –, but also in hominid evolution.

References

Bloom, Paul (1994). Generativity within language and other cognitive domains. *Cognition, 51,* 177-189

Butterworth, Brian (1999). *The mathematical brain*. London: Macmillan.

Deacon, Terrence W. (1997). *The symbolic species. The co-evolution of language and the brain.* New York: Norton & Co.

Dehaene, Stanislas (1997). *The number sense: How the mind creates mathematics.* Oxford University Press.

Dehaene, Stanislas; Dehaene-Lambertz, Ghislaine, & Cohen, Laurent (1998). Abstract representations of numbers in the animal and human brain. *Trends in Neurosciences, 21 (8),* 355-361.

Frege, Gottlob (1884). *Die Grundlagen der Arithmetik. Eine logisch mathematische Untersuchung über den Begriff der Zahl.* Breslau: Wilhelm Koebner. [English transl. by J. L. Austin. Oxford: Blackwell 1950].

Fuson, Karen C., & Hall, James W. (1983). The acquisition of early number word meanings: A conceptual analysis and review. In H. P. Ginsburg (Ed.), *The development of mathematical thinking.* New York: Academic Press.

Ho, Connie Suk Han, & Fuson, Karen C. (1998). Children's knowledge of teen quantities as tens and ones: Comparisons of Chinese, British, and American kindergartners. *Journal of Educational Psychology, 90 (3)*, 536-544.

Hurford, James, R. (1987). *Language and number: The emergence of a cognitive system.* Oxford: Blackwell.

Jackendoff, Ray (1999). Possible stages in the evolution of the language capacity. *Trends in the Cognitive Sciences 3 (7),* 272-279.

Krantz, David H.; Luce, R. Duncan; Suppes, Patrick, & Tversky, Amos (1971). *Foundations of measurement*, 3 vols. New York: Academic Press.

Locke, John (1690). *Essay concerning human understanding*. Oxford: Clarendon Press, 1975.

Miura, Irene T.; Okamoto, Yukari; Kim, Chungsoon; Steere, Marcia, & Fayol, Michel (1993). First graders' cognitive representation of number and understanding of place value: Cross-national comparisons: France, Japan, Korea, Sweden, and the United States. *Journal of Educational Psychology, 85 (1),* 24-30.

Narens, Louis (1985). *Abstract measurement theory.* Cambridge, Mass.: MIT Press.

Pinker Steven (1994). *The language instinct*. New York: Morrow.

Roberts, Fred S. (1979). *Measurement theory*. Reading: Addison-Wesley.

Wiese, Heike (1997). *Zahl und Numerale. Eine Untersuchung zur Korrelation konzeptueller und sprachlicher Strukturen.* Berlin: Akademie-Verlag.

Wiese, Heike (2001). *Numbers, language and the human mind* (in preparation).

Wynn, Karen (1998). Numerical competence in infants. In Donlan, Chris (Ed.), *The development of mathematical skills.* Hove: Psychology Press.

Selection Procedures for Module Discovery: Exploring Evolutionary Algorithms for Cognitive Science

Janet Wiles (*j.wiles@csee.uq.edu.au*)
Ruth Schulz (*ruth@csee.uq.edu.au*)
Scott Bolland (*scottb@csee.uq.edu.au*)
Bradley Tonkes (*btonkes@csee.uq.edu.au*)
Jennifer Hallinan (*J.Hallinan@imb.uq.edu.au*)
The University of Queensland, Brisbane QLD 4072 Australia

Abstract

Evolutionary algorithms are playing an increasingly important role as search methods in cognitive science domains. In this study, methodological issues in the use of evolutionary algorithms were investigated via simulations in which procedures were systematically varied to modify the selection pressures on populations of evolving agents. Traditional roulette wheel, tournament, and variations of these selection algorithms were compared on the "needle-in-a-haystack" problem developed by Hinton and Nowlan in their 1987 study of the Baldwin effect. The task is an important one for cognitive science, as it demonstrates the power of learning as a local search technique in smoothing a fitness landscape that lacks gradient information. One aspect that has continued to foster interest in the problem is the observation of residual learning ability in simulated populations even after long periods of time.

Effective evolutionary algorithms balance their search effort between broad exploration of the search space and in-depth exploitation of promising solutions already found. Issues discussed include the differential effects of rank and proportional selection, the tradeoff between migration of populations towards good solutions and maintenance of diversity, and the development of measures that illustrate how each selection algorithm affects the search process over generations. We show that both roulette wheel and tournament algorithms can be modified to appropriately balance search between exploration and exploitation, and effectively eliminate residual learning in this problem.

Introduction: EC and Cognitive Science

Evolutionary computation (EC) is increasingly used in cognitive science, both for evolving cognitive models and for modeling evolutionary processes.

Many algorithms use evolutionary search in one form or another. No single search algorithm will be optimal for all tasks (a thesis colloquially known as "no free lunch", Wolpert & Macready, 1996). In any simulation study, characteristics of the task need to be taken into account in the selection of algorithms. However, to many cognitive science researchers it is not clear which aspects of tasks are important in the design of a search process, and what properties of evolutionary search algorithms need to be taken into account to select an appropriate design.

This study is part of a wider program of research whose goal is to enhance the effective use of evolutionary computation techniques in cognitive science domains. This program involves assessing the performance of popular evolutionary algorithms on tasks of interest to cognitive scientists.

Current areas in cognitive science that are utilizing EC methods include the direct modeling of evolutionary processes, such as the role of learning in evolution, learning as a local search technique in a genetic algorithm, the evolution of modularity, the evolution of cooperation, and the evolution and learnability of language (e.g., see the biennial "Evolution of Language" conferences, or the Evolutionary Computation "Special Issue on EC and Cognitive Science", Wiles & Hallinan, 2001).

Other domains use evolutionary algorithms for optimization, for example, testing theories of infant development; modeling populations of individuals engaged in cognitive tasks; testing outcomes following damage in neural network models; and exploring the range of behaviors in a dynamic model of an artificial language learning task.

In all of the cognitive science domains mentioned, evolutionary algorithms have been tested on specific problems, but little work has been done at a methodological level to characterize the nature of the tasks per se, and the way in which they interact with the evolutionary algorithms chosen. Many factors affect the performance of evolutionary algorithms, including the choice of fitness function, representation of the genome, population size, selection technique, and genetic operators.

Learning and EC

For this study, the area of interest is the interaction of learning and evolution known as the Baldwin effect, first formalized as a computational problem by Hinton and Nowlan (1987). Hinton and Nowlan's simulation of the Baldwin effect provided insight into how learning can guide evolution within a Darwinian, rather than a Lamarkian evolutionary framework.

In Hinton and Nowlan's model, each individual consisted of a bit string representing a simple neural network with twenty connections, which must be set correctly via either learning or evolution. A network that achieves the correct settings has a fitness dependant upon the time required to achieve the correct settings, while all incorrect networks have equal, minimal fitness. This task has a single fitness peak, which is surrounded by a perfectly flat fitness landscape, making it a classic needle-in-a-haystack problem (henceforth referred to as the *haystack* problem). The task is analogous to finding the components of a module in which no partial credit is given for partial solutions. Issues of modular design were popularized by Dawkins in the Blind Watchmaker (Dawkins, 1986), and are particularly relevant to understanding the evolution of complex cognitive systems.

The haystack task requires exhaustive search if genetic operators alone are used (crossover and mutation). However, if each agent modeled in the search population is allowed to perform some local searching, then the task can be solved by a much smaller population.

Hinton and Nowlan used a population size set to approximately the square root of the size of the search space, with each agent able to search on average a portion of the search space also equal to the square root of the size of the space. The choice of population size and local search space balanced the need for a population to have sufficient diversity to cover the space, and sufficient flexibility to find the "needle" (maximum fitness) in almost every trial.

Computationally, each individual is implemented as a string of twenty "genes", each of which may be either 1, 0, or ? (question mark). The ? represents a learnable gene. The individual learns by guessing 0 or 1 with a probability of 0.5. The target pattern is a string of twenty 1s. The number of guesses required to achieve this target is recorded and used to calculate the individual's fitness. The next generation is created by repeatedly selecting two parents, to produce pairs of new individuals. Parents are probabilistically selected proportional to the individual's fitness relative to the total population fitness.

Hinton and Nowlan (1987) demonstrated that under these conditions, the ability to learn, represented by ?s, was replaced by appropriate instincts, represented by 1s. The number of 1s rose from an initial 25% of alleles in the population to 50-80% after 50 generations, with the remainder of the alleles ?s. Non-target alleles, represented by 0s, disappeared from the population.

An interesting feature of Hinton and Nowlan's simulation is the persistence of learnable genes in the population once it has stabilized. Hinton and Nowlan suggested that there is very little selective pressure in favor of genetically specifying the last few connections, since learning can occur in very few guesses. Several researchers have studied the phenomena of the residual question marks in the haystack problem and identified a variety of factors, including selection pressure and drift as significant factors in the results (Belew, 1989, Harvey, 1993).

In a previous study, we analyzed multiple simulations of the haystack problem to identify the characteristics of two classic selection algorithms (one fitness proportional and the other rank based) with respect to exploitation and exploration of the fitness landscape (Wiles et al., in press). These simulations demonstrated that fitness proportional selection finds good solutions and the average fitness of a population rises quickly, but at high fitness levels the population drifts gradually to homogeneity (all the alleles in one position on the chromosome are identical for all individuals in a population). Residual learning is frequently a result of an interaction between a pseudo-founder effect (dominance by one early successful solution) and drift to homogeneity at one or more of the genes. Selection by rank has the opposite effect, with populations drifting initially, until a critical mass find good solutions (or until an allele becomes homogeneous in 0s, resulting in an unsuccessful trial). Of the successful trials, at high fitness levels, populations converge to homogeneity based on fitness, rather than drift. By comparing fitness level and number of homogeneous genes to generation number, the relative effects of drift and selection pressure can be monitored during evolution.

The analyses in our previous study provide tools to understand how selection pressures are working during trials. The two techniques produce very different characteristic performance. Fitness proportional selection has initially fast fitness increases followed by slow convergence, whereas rank-based selection has initially slow and erratic fitness increases followed by fast convergence.

For the haystack problem, neither of these selection methods can be considered optimal in balancing the exploration of possible solutions with the exploitation of good solutions. Fitness proportional selection has too strong an exploitation of early successful solutions, leading to a pseudo-founder effect, and insufficient pressure to optimize when most of the population have good solutions. In contrast, rank-based selection has insufficient exploitation of its good early solutions, allowing drift to reduce the diversity of alleles available before fitness pressures shape the search space.

Method

In this study, we report three sets of simulations. The first set replicates our previous work on the classic fitness proportional (roulette wheel) and stochastic

rank-based (tournament) methods and is reproduced here for comparison.

The second set of simulations was designed to investigate other selection algorithms on the haystack task, and also to test whether the analysis would be useful for their evaluation. For this set, we designed two additional algorithms to combine the search characteristics of fitness proportional and rank-based selection. The first operator was designed to exhibit fast fitness rises and fast convergence, and the second to exhibit slow fitness rises and slow convergence.

The third set of simulations directly addressed the problems inherent in fitness proportional and rank-based selection using modifications suggested in the literature. To modify fitness proportional selection, the expected number of offspring for any one individual was scaled in proportion to its deviation from the mean fitnesses of other individuals, which balances the selection pressure over a trial. To modify rank-based selection, the two fittest individuals (replacing the offspring of one pair of parents) were copied to the next generation, thus preserving good solutions once found. In the next section, the algorithms are described, and then the results summarized and presented together for ease of comparison.

Simulation details for the haystack problem

Hinton and Nowlan (1987) modelled the Baldwin Effect using a simple genetic algorithm, with no mutation and a crossover value of 1.0; each pair of parents undergoes crossover once during each reproduction event. The next generation is created by repeatedly selecting two parents for each pair of new individuals. The probability of selecting an individual as a parent is proportional to its fitness divided by the total population fitness. The fitness, *f*, of an individual is calculated using the recorded number of guesses, *g*, taken to find the target:

$$f = 1 + \frac{(L-1)(G-g)}{G} \qquad (1)$$

where G is the maximum number of guesses allowed and L is the length of the chromosome. In Hinton and Nowlan's model, $G = 1000$, $L = 20$, and the population size, $N = 1000$.

We implemented Hinton and Nowlan's model, and as in our previous work, selection of each parent was implemented using a fitness proportional algorithm. After selecting two parents, a crossover point was chosen at random, and two new individuals were then created. Parameters were set similar to Hinton and Nowlan (1987), with initial proportions of 1, 0 and ? alleles set to 0.25, 0.25 and 0.50 respectively. A minor change from their parameters was setting both population size and maximum number of guesses to 1024 instead of 1000. All trials were run until the population converged (homogeneous in all genes).

For each selection method, 100 trials were performed. We report the proportion of trials that successfully eliminated all zeros from the population, the average number of residual question marks at the end of each trial, and the average number of generations to homogeneity (see Table 1). The average fitness pressures in the early and late stages of trials were calculated as the average number of generations until the fitness rose to 50% of the maximum (Stage 1) and from the midpoint to final convergence (Stage 2). By defining these values, the relative selection pressures early and late in a run can be compared. The average fitness of the population when the first gene in each trial became homogeneous was also calculated (see Table 1, column 4). This measure shows the potential exploration available to the algorithm.

Set 1: Original algorithms

Traditional roulette wheel (fitness proportional) selection: The fitness of an individual is determined using equation (1) given above, and the selection procedure for two parents is as described for Hinton and Nowlan's simulations.

Tournament (rank based) selection: In this algorithm two candidates are selected at random from the parent population, and the individual with the higher fitness becomes a parent. The probability of being selected as a parent for the next generation therefore depends on the relative rank of an individual within the population, rather than its proportional fitness. Under tournament selection, the reduced fitness differential later in evolution does not change the ranking of individuals and selection pressure is maintained as long as there are different fitnesses within the population.

Set 2: Modified algorithms

Roulette with ranking: In order to produce a selection strategy that should both begin and end rapidly, ranking was added to roulette wheel selection. This algorithm has also been called stochastic tournament (attributed to Wetzel by Goldberg, 1989). Continued pressure after the initial fast start means that selection will force the population to converge, rather than simply drifting to homogeneity.

To select each parent, two candidates are chosen using roulette wheel selection. The fittest of these two individuals becomes one parent, as in tournament selection. A second parent is selected in the same way. The fitness-proportional selection of candidates enables very successful individuals to have many offspring, in a similar manner to roulette wheel selection. The addition of a tournament between two candidates ensures that as fitness differentials decrease later in trials, the selection

pressures continue. The strategy is identical in all other ways to the others that have been used previously.

Probabilistic tournament: The second variation is a strategy that is designed to start slowly and end slowly. For this strategy, tournament selection was modified to include the proportional elements of roulette wheel strategy.

For each parent, two candidates are chosen randomly from the parent population. The one that will become a parent is chosen using proportional selection based on the fitness of these two individuals. That is, the one that is fitter will be more likely to be chosen than the less fit individual, but both still have a chance of being a parent. The selection of candidates with equal probability means that each individual, even the fittest one, can expect on average to contribute genes to a maximum of four offspring. The second parent is chosen in the same way, and reproduction continues as in the other selection strategies.

Set 3: Optimized algorithms

Sigma Scaled Roulette: Amongst the known problems with roulette wheel selection is the variable selection pressure between early and late stages in a trial and the premature convergence of populations with inadequate exploration of the search space (Mitchell, 1996). A variety of modifications of roulette wheel selection have been proposed. One such mechanism is to balance the selection pressures evenly throughout a trial. Sigma scaled roulette is a renormalized version of roulette wheel. We use the description given by Mitchell (1986, who credits an early unpublished manuscript of Forrest from 1985). The expected number of offspring, *E*, is calculated from the mean and standard deviation of the fitnesses of the population:

$$E = 1 + \frac{f(i) - m}{2\sigma} \qquad if \sigma \neq 0$$

$$= 1 \qquad if \sigma = 0$$

where $f(i)$ is the fitness of individual i, m is the mean fitness of the population and σ is the standard deviation. This means that an individual with a fitness equal to the mean will gain a slice of the roulette wheel proportional to one unit. An individual with fitness one standard deviation above the mean will (on average) gain a slice proportional to 1.5 units, and one with fitness one standard deviation below will gain a slice proportional to 0.5 units. If the expected value for an individual is less than 0.1, then the slice is set to 0.1. The total size of the wheel is the sum of the slices of all individuals in the population. The expected number of offspring is proportional to the size of the slice, with corrections for the very small slices. For each pair of parents selected, two offspring are produced. Using the standard deviation of the fitness maintains a constant selection pressure in the population throughout a trial.

Elite tournament: The slow initial period of all trials during tournament selection is a known problem. Even when a good solution is found, recombination of parents results in disruption of the solution and drift (rather than selection) can lead to homogeneity in one or more of the genes. Many researchers use elitism to preserve good solutions (first introduced in the 1970s by de Jong, according to Mitchell, 1996). In this strategy, one or more individuals with the highest fitnesses are copied to the next generation unchanged. In our implementation, elite tournament is identical to tournament selection, except that the individuals with the two highest fitnesses are copied to the next generation.

Results and Discussion

The results from all three sets of simulations have been collated in Table 1, which shows the performance of each selection technique, the average number of residual question marks, and the number of generations to convergence (i.e., all genes become homogeneous) over 100 trials. Figures 1-3 show the selection pressures during early and late stages over 10 trials. Stage 1 is the average number of generations taken to reach a fitness value of 10 (i.e., 50% of the maximum fitness), the point at which fitness increases most rapidly in this task. Low values (few generations) indicate high initial selection pressure and high values indicate low selection pressure. Stage 2 is the average number of generations from this level of fitness to convergence of the population and indicates the selection pressure after the initial increase in fitness. The mean population fitness when the first allele becomes homogeneous generally indicates how good the solution will be. If an allele is homogeneous before the population is very fit, the solution tends to be poor.

The simulations in Set 1 (roulette wheel and tournament) provide a benchmark for the later studies. Performance in these simulations was consistent with our previous results (Wiles et al. in press). Using roulette wheel selection, all trials eliminated zeros from the population, and at convergence, individuals had an average of 1.4 residual question marks (stdev. 0.9). The change in selection pressure is revealed by the average number of generations spent in Stage 1 (10) versus Stage 2 (1437). See Table 1 for the means and standard deviations of all results.

Using tournament selection, 85/100 trials eliminated all zeros. In the successful trials, individuals retained an average of 1.2 residual question marks, with a much higher variance (stdev. 2.2). The average time spent in Stages 1 and 2 is reversed in this case, with 185 generations in Stage 1 and 23 generations in Stage 2.

The interaction between Stages 1 and 2 for roulette wheel and tournament selection is clear in Figure 1.

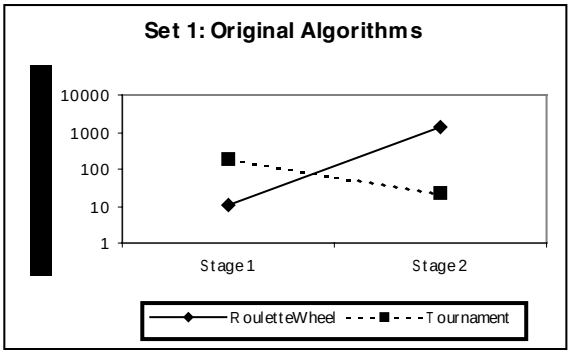


Figure 1. Time to convergence in Set 1, the original roulette wheel and tournament selection algorithms. Note that the y-axis is logarithmic. The algorithms show clear differences in behaviour, with roulette wheel faster in Stage 1, and tournament faster in Stage 2.

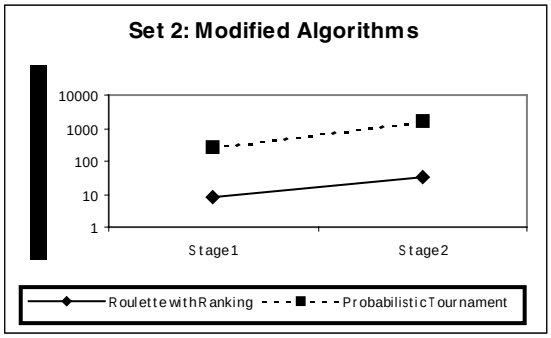


Figure 2. Time to convergence in Set 2, the modified algorithms. The Stages 1 and 2 components from Set 1 (see Fig 1), have been recombined as intended, to produce one algorithm in which both Stages 1 and 2 are fast (roulette with ranking), and the other algorithm in which both Stages 1 and 2 are slow (probabilistic tournament). Note that neither of these algorithms eliminate residual question marks.

In Set 2, the number of successful trials and residual question marks are similar to those from Set 1, but the time spent in Stages 1 and 2 differed markedly, as expected. Roulette wheel with ranking was fast in both stages (averages 7.6 and 33 generations respectively), and stochastic tournament was slow in both stages (average 249 and 1624 generations respectively, see Figure 2).

In Set 3, the original roulette and tournament selection procedures were modified to address their major known weaknesses, and both showed considerable improvement in optimization performance (as evidenced by the number of residual question marks). All trials eliminated zeros from the population, the time to convergence was short and very few residual question marks remained (an average of 0.02 in sigma scaled roulette and 0.06 in elite tournament).

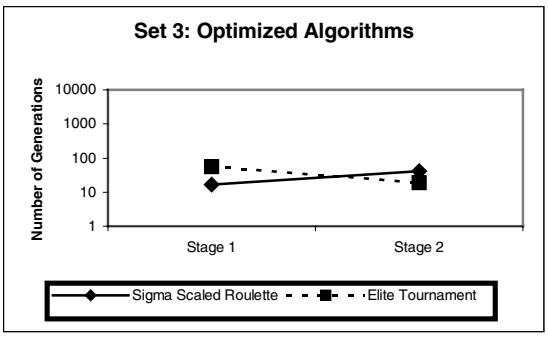


Figure 3. Time to convergence in Set 3, the optimized algorithms. Both sigma scaled roulette and elite tournament eliminate virtually all residual question marks and are much faster than the algorithms in Sets 1 and 2 (cf. Figs 1 and 2).

The average time spent in Stage 1 and Stage 2 was also much more balanced (16 and 42 generations respectively in sigma scaled roulette and 56 and 19 generations respectively in elite tournament, see Figure 3).

Premature convergence is a known problem for these evolutionary algorithms. Tracking progress towards homogeneity can therefore provide valuable information. The average fitness at which the first gene becomes homogeneous provides a quantifiable measure of diversity at a significant point in a simulation. This fitness value was recorded for each selection regime (see Table 1, column 5). Higher values (maximum is 20) indicate that higher levels of diversity are maintained in the population. For problems in which hitchhiking genes (sub-optimal genes that are carried by pseudo-founders in a population) are liable to cause problems such as in the haystack problem, this measure is a good indicator of potential problems with premature convergence. The tournament-based algorithms that have trials that do not eliminate all zeros show the lowest values with average population fitnesses at the first homogeneous alleles of 9.1 and 6.0 for Sets 1 and 2 respectively. Values for the corresponding roulette wheel-based algorithms are higher (16.7 and 13.6 respectively), but are not optimal. The best algorithms, those in Set 3 have the highest values (19.3 for sigma-scaled roulette and 19.9 for elite tournament) indicating that none of the trials suffered from premature convergence.

The combination of relatively balanced fitness pressures in Stages 1 and 2, short times to convergence, and high fitness before diversity is reduced indicate that both selection algorithms in Set 3 are well-adapted to the haystack task.

Conclusions

One specific conclusion from these experiments is that residual learning is not an inherent aspect of the Baldwin effect. Rather, it is a consequence of the way

the fitness landscape is searched, and the application of selection pressures at different stages. The methodological studies presented in this paper are one way to explore such issues. Further work is needed to tie these results to biologically-plausible learning scenarios, but that is beyond the scope of this study.

At a more general level, the simulations show that the haystack task is one for which tailoring of the algorithm makes a qualitative difference to the behaviors observed. Specific issues addressed in this study concern the characteristics of the algorithms and the nature of the landscape.

The simulations of the original algorithms illustrate properties such as premature convergence in roulette wheel and the dangers of early homogeneity in one or more genes due to drift in tournament selection. With appropriate modifications, the optimized algorithms achieve a balance between exploration and exploitation, resulting in convergence to good solutions. Residual learning can be almost eliminated, and performance on the haystack problem can be near optimal.

These results illustrate the need for a characterization of task types in cognitive science, and a characterization of evolutionary algorithms and their performance on these tasks. Such a classification would facilitate the tailoring of algorithms to particular problems, and has the potential to significantly reduce artifacts due to implementation details.

Acknowledgements

This project was supported by a CSEE summer grant to RS, and an APA to SB.

References

Baldwin, J. M. (1896). A new factor in evolution. *American Naturalist* 30: 441 – 451. Reproduced in (eds.) Belew, R.K. & Mitchell, M.,. *Adaptive Individuals in Evolving Populations*, *Proceedings Volume XXVI, Santa Fe Institute Studies in the Sciences of Complexity*. Addison-Wesley, Reading, MA.

Belew, R. K. (1990). Evolution, learning and culture: Computational metaphors for adaptive search. *Complex Systems* 4 (1), 11-49.

Dawkins, R. (1986). *The Blind Watchmaker: Why the evidence of evolution reveals a universe without design*, NY: Norton.

Goldberg, D.E. (1989), *Genetic algorithms in search, optimization, and machine learning*, Addison-Wesley, Reading, MA.

Harvey, I. (1993). The puzzle of the persistent question marks: A case study of genetic drift. *Computer Science Research Paper Serial No. CSRP 278*, The University of Sussex. (Also published in S. Forrest, (Ed.) *Proceedings of the Fifth Int. Conf. on Genetic Algorithms*, Morgan Kaufmann, 1993.)

Hinton, G. E. & Nowlan, S. J. (1987). How learning can guide evolution. *Complex Systems* 1: 495 –502.

Maynard Smith, J., (1998). *Evolutionary Genetics*, second edition. Oxford University Press: NY.

Mitchell, M. (1996). *An Introduction to Genetic Algorithms*. MIT Press: Cambridge, MA.

Wiles, J. and Hallinan, J.S. (eds) *IEEE Trans. on Evolutionary Computation Special Issue on EC and Cognitive Science.* 5 (2), 2001.

Wiles, J., Schulz, R., Hallinan, J., Bolland, S., and Tonkes, B., *Probing the Persistent Question Marks*, to appear in the Proceedings of GECCO, 2001.

Wolpert, D.H. and Macready, W.G. (1997), No Free Lunch Theorems for Optimization. *IEEE Trans Evolutionary Computation*, April 1997, pp. 67-82.

Table 1. Summary of Numerical Results

Selection Strategy	All Trials	Successful Populations		
	Proportion of trials that eliminated 0s	Residual homogeneous ?s [Mean (SD) of 20 trials]	Generations to homogeneity [Mean (SD)]	Av fitness at 1st homogeneous allele [Mean (SD) of 10 trials]
Roulette Wheel	100	1.44 (0.91)	1448.16 (734.66)	16.70 (1.68)
Tournament	85	1.24 (2.22)	208.67 (142.60)	9.13 (8.81)
Roulette with Ranking	100	1.27 (1.25)	41.45 (9.58)	13.62 (3.08)
Probabilistic Tournamt	80	2.08 (1.79)	1873.66 (1035.19)	6.01 (8.19)
Sigma Scaled Roulette	100	0.02 (0.14)	57.93 (5.41)	19.34 (0.53)
Elite Tournament	100	0.06 (0.34)	75.01 (20.12)	19.90 (0.08)

How learning can guide evolution in hierarchical modular tasks

Janet Wiles (janetw@csee.uq.edu.au)
School of Psychology and School of Computer Science and Electrical Engineering
University of Queensland, Qld 4072 Australia
Bradley Tonkes (btonkes@csee.uq.edu.au)
James R. Watson (jwatson@csee.uq.edu.au)
School of Computer Science and Electrical Engineering
University of Queensland, Qld 4072 Australia

Abstract

This paper addresses the problem of how and when learning is an aid to evolutionary search in hierarchical modular tasks. It brings together two major areas of research in evolutionary computation, the performance of evolutionary algorithms on hierarchical modular tasks, and the role of learning in evolutionary search, known as the Baldwin effect. A new task called the jester's cap is proposed, formed by adding learning to Watson, Hornby and Pollack's Hierarchical-If-and-only-If, function, using the simple guessing framework of Hinton and Nowlan's Baldwin effect simulations. Whereas Hinton and Nowlan used a task with a single fitness peak, ideally suited to learning, the jester's cap is a hierarchical task that has two major fitness peaks and multiple sub-peaks. We conducted a series of simulations to explore the effect of different amounts of learning on the jester's cap. The simulations demonstrate that learning aids evolution only in search spaces in which the simplest level of modules are difficult to find. The learning mechanism explores local regions of the search space, while crossover explores neighborhoods in higher-order modular spaces.

Introduction

This paper addresses the problem of how and when learning is an aid to evolutionary search in hierarchical modular tasks. It brings together two major areas of research in evolutionary computation (EC), the performance of evolutionary algorithms (EAs) on hierarchical modular tasks, and computational models of the role of learning in evolutionary search, known as the Baldwin effect.

We begin with a brief review of modular tasks that have been proposed to explore the performance of evolutionary algorithms, and then briefly describe the Baldwin effect. We then describe a specific task, the *jester's cap*, that incorporates learning into a hierarchical modular task. In many simulation tasks, learning is costly and does not improve the performance of an evolutionary algorithm (French and Messinger, 1994; Mayley, 1996; Pereira et al., 2000). This study is as much an investigation of things that don't learn, as of ones that do.

There are many types of EAs, and the field of evolutionary computation is still determining features of problems that are easy or hard for a particular class of EA, and the conditions under which such algorithms will perform better than other search techniques. In evolutionary computation, characterization of an EA's performance concerns not just optimization per se, but the behaviors of populations as a whole, reflecting their original motivations as models (albeit abstract ones) of real evolutionary processes.

Some of the oldest and most popular techniques for evolutionary search are genetic algorithms (GAs), which use crossover as their major search technique. Originally developed by Holland (1992), their efficacy is thought to be based on groups of genes acting together as modules (or *building blocks*, to use Holland's original terminology), and have been studied extensively since (for general introductions see Goldberg, 1989 and Mitchell, 1996).

A variety of modular tasks have been proposed to study the conditions under which GAs outperform comparable search techniques. The most widely known of these are the Royal Road (RR) problems introduced by Mitchell et al. (1992). However, some forms of hill-climbers were found to easily outperform the GA, and a variety of tasks that incorporate deceptive elements have been defined (s.a., RR4 by Mitchell et al., 1994; HDF by Pelikan and Goldberg, 2000; hdf by Holland, 2000).

An alternative approach to incorporating deceptive elements is to define a fitness function with two or more conflicting maxima. Watson et al. (1998) defined Hierarchical-If-and-only-If (*H-IFF*) as such a function. H-IFF is a simple function that is hierarchical, modular, is not searchable by mutation, but is amenable to search by crossover. Its defining characteristics are two fitness peaks at opposing ends of the search space. Combinations of the sub-components that comprise each level of the competing hierarchies cause many sub-optimal peaks and consequently many local minima (see below for the complete definition).

Before proceeding further with the computational aspects, it is worthwhile considering the relevance of module building to many areas of cognitive science. The role of modules in evolution has long been recognized (e.g., Dawkins, 1986). In evolutionary psychology there is a particularly strong interest in modules, in part due to Tooby and Cosmides (1994) claims that humans have behavioral modules analogous to other mental functions. By studying building block problems, we are considering the types of processes that allow species to evolve varieties of modules, and their combination into complex mental organs. For example, echolocation in bats requires both the ability to emit and to receive high fre-

quency sounds. Each of these abilities has utility as a module in its own right, but the combination provides an additional capacity that goes beyond the cumulative benefit of the independent components.

The Baldwin effect: How learning can guide evolution

Under Darwinian inheritance, the things that an animal learns during its life cannot be passed directly to its offspring via the genotype. However, researchers in the late 19th century (Baldwin, 1896; Morgan, 1896), proposed a way in which learned behaviors could be incorporated into a genome over many generations (i.e., become instinctual). The mechanism is purely Darwinian, and relies on gradual changes in the distribution of genes in a population, as the following rationale explains.

Consider a population of agents comprising a variety of search strategies with initially random starting points and a range of search radii. The starting point and search radius of an agent is its "bias". An evolutionary algorithm that selected for speed in finding a point in the search landscape over many generations would evolve a population of individuals that had starting points close to the fitness maximum and small search radii. That is, behaviors that were initially interpreted as general learning abilities would, over time, become innate. No information about the content of learning is passed from parent to offspring, but in the general variation across the population, some individuals would by chance have starting points closer to the fitness maximum and smaller search radii (i.e., slightly stronger biases). These individuals would have more offspring than those with weaker biases, and in environments with fixed fitness functions, innate behaviors would gradually replace learned ones.

This process is often called the Baldwin effect and has two interesting components. The first is the explanation of how learned behaviors can become innate as described above. The other is the power of learning to augment genetic search to build complex modules. Agents that can search their local environment will be able to explore whole regions of search space in their lifetimes, rather than the single point of their own genotype. In this way, learning can enable an evolutionary algorithm to solve problems that are too costly for genetic search alone.

The first computational simulations of the Baldwin effect were by Hinton and Nowlan (1987). They used a needle-in-a-haystack (NIAH) problem, in which the maximum fitness of an agent corresponded to a genotype comprising all ones. Each gene could be one, zero or question mark. The ones and zeroes were fixed values that did not change during an individual's lifetime. The question marks were learnable genes, which could change during a lifetime. Hinton and Nowlan showed that the zero alleles quickly dropped out of the population and the number of question marks reduced over time. Hinton and Nowlan's study is a landmark in EC because it was the first computational demonstration showing how learning can guide evolution.

Hinton and Nowlan's original simulations have stimulated a considerable body of literature, which is only briefly mentioned here: Belew (1990) replicated and extended the original study, adding changing environments and cultural advantage; Harvey (1993) showed how remnants of residual learning are due to genetic drift; French and Messinger (1994) investigated under what conditions learning supplements genetic search; Mayley (1996) demonstrated how learning is first selected for, then against as evolution progresses; and Mitchell and Belew (1996) discuss issues arising from the original study. A useful reference is the edited volume of papers relating to learning and evolution by Mitchell and Belew (1996), which includes reprints of the original papers by Baldwin (1896), Morgan (1896), and Hinton and Nowlan (1987) as well as many other related studies.

Learning in hierarchical modular spaces

The issues in this study follows from French and Messinger's (1994): Consideration of the circumstances under which is it reasonable to expect that learned behaviors will firstly enhance evolutionary search, and secondly be gradually replaced by genetically specified traits. However, we take an alternative approach and investigate whether learning aids evolution in search spaces that contain competing modules. One way of thinking about this issue is in terms of the difficulty of the search task compared to the operators that are available. In Hinton and Nowlan's NIAH task, the set of twenty ones can be considered as one module, with no intermediate fitness levels for partial results. The search task for the genome is too large to find by populations of size substantially smaller than that of the search space. In this case, learning performs the function of a local search through points close together in Hamming distance. The local search supplements the genetic search, effectively smoothing the search landscape (see top, Figure 1).

The NIAH task is a particularly pathological as it contains no partial information to guide a search process. The majority of tasks of interest in EC and cognitive science have some internal structure, or distinct modules. The most tractable problems that have modular structure are those in which the genes that comprise modules can be selected independently of the settings or global structure of other genes.

As described above, a variety of such tasks have been proposed to explore the functionality of GAs, including the Royal Road problems (Forrest and Mitchell, 1993; Mitchell et al., 1994). The Royal Road problems had only one fitness peak, and hence hill climbing strategies worked well (see Mitchell, 1996 for a summary of the reasons).

The H-IFF problem has the interesting property of symmetry around diametrically opposed fitness peaks with many sub-optimal peaks and consequently many local minima (see bottom, Figure 1).

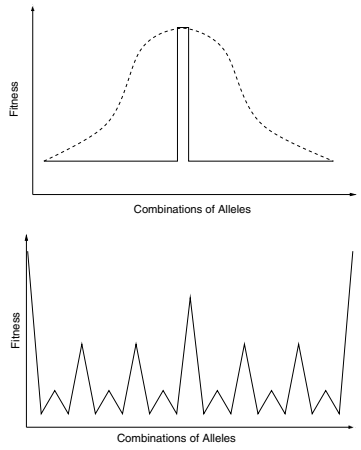


Figure 1: Fitness landscapes in different search tasks. (top) The needle-in-a-haystack task has a single fitness peak, and learning smoothes the search landscape around the peak (adapted from Hinton and Nowlan, 1987, Figure 1). (bottom) A slice through the fitness landscape of H-IFF, showing the multiple fitness peaks and the two maxima at all ones and all zeros (adapted from Watson and Pollack, 1999, Figure 1).

Tasks: H-IFF and the Jester's Cap

H-IFF (Watson and Pollack, 1999) is a function defined on bit-strings of length 2^n. The fitness value of a particular string is defined in terms of hierarchical 'building blocks' which are sub-strings of the main bit-string. The building block at the highest level of the hierarchy is the entire bit-string (i.e., all 2^n bits). Each building block is recursively divided into two equally-sized blocks, except for blocks of size one, which cannot be divided. For a building block to be correctly set, it must consist of either all 1s or all 0s. The value of a correctly set building block of size n is 2^n *plus* the sum of the values of its two sub-blocks (whose values depend on the sub-sub-blocks). Thus, the overall value of a bit-string of length 2^n is the sum of values for the building blocks of sizes $1, 2, 4, \ldots, 2^n$. The optimum bit-strings consist of either all 0s or all 1s so that they are rewarded for building blocks of every size. In the simulations in this paper, we use $2^n = 32$. The evaluation of the H-IFF function is more easily understood by way of example, shown for an 8-bit string in Figure 2.

As described above, the major difference between H-IFF and the more well known Royal Road (RR) function is that RR has a single optimal bit-string (all 1s) and significantly, *no local optima* other than the global optimum (although there are local plateaus). By comparison, H-IFF has two optimal bit-strings and, for bit-strings of length $l = 2^n$, there are $2^{l/2}$ local optima.

In this paper, we apply the learning-based approach of Hinton and Nowlan's simulations to the H-IFF function. We call this modified version the jester's cap. Specifically, we consider a genome of 32 genes, each of which may be a 0, 1 or ?. During its lifetime, an individual tries to 'learn' the best setting of the ? genes. Each of the ? genes can be set to either 0 or 1, and the resulting bit-string, comprising all 1s and 0s is evaluated with the H-IFF fitness function, described above. We take a very simplistic view of learning (as in Hinton and Nowlan's original simulations), and give the agent N attempts at guessing the best setting. The agent tests N replacements of all of the question marks with random choices of 1s and 0s. After N guesses, the best guess (i.e., that which maximizes fitness) is taken and the fitness of the agent is the H-IFF fitness of that guess. (Unlike Hinton and Nowlan's simulations, there is no scaling of the fitness based on the number of guesses required.) For example, an agent with the genome 0??1 may generate the guesses 0101, 0111 and 0011 which evaluate to 4, 6 and 8 respectively. The highest scoring guess (0011) is taken as the 'learned' setting. However, this 'learned' setting is not passed on during reproduction, it is the initial genome, 0??1 that is used in reproduction.

0	**0**	**1**	**0**	**1**	**1**	**1**	**1**	Value
1	1	1	1	1	1	1	1	8
2		—		2		2		6
—				4				4
—								0

Figure 2: Evaluation of H-IFF for the bit-string 00101111 showing hierarchical decomposition. For this bit-string, H-IFF evaluates to $8+6+4+0 = 18$. Note that the maximum obtainable value for each level of the hierarchy is 8, so the maximum value for H-IFF on bit-strings of length 8 is 32. In general, there will be $n+1$ levels of building blocks. Within these levels there will be $2^n/k$ building blocks of size k, each of which has value k. The optimal bit-strings of length 2^n therefore have value $(n+1).2^k$. The minimum value for H-IFF on bit-strings of length $l = 2^n$ is l. Such a bit-string contains all building blocks of size 1 but no higher-level blocks.

Simulation 1: The Jester's Cap

We consider the jester's cap task with three variations of the amount of learning time available to the agents: no learning (replicating the H-IFF task), a small amount of learning ($N = 10$) and a moderate amount of learning ($N = 100$). A population comprising 500 individuals is embedded within a genetic algorithm. In this initial population, 50% of the genes are ?s, 25% are 1s and 25% are 0s, except in the case without learning, where there are no ?s and equal proportions of 1s and 0s. In each generation, the fitness of each of the agents is determined, after learning when appropriate. These fitness values are used to determine the parents for the next generation, those agents with higher fitness being (probabilistically) more likely to be selected as parents than those with lower fit-

nesses. (We used a sigma-scaled roulette algorithm for choosing the parents, see Wiles et al. for further details.) Each pair of parents is used to generate two new offspring using single point crossover (zero mutation). In this recombination technique a 'cut-point' is selected at a random position on the genome. One offspring is generated by joining the genes to the left of the cut-point in parent 1 to the genes to the right of the cut-point in parent 2. The second offspring is formed by the reverse combination.

The idea behind this evolutionary approach is that in the initial population some agents will, by chance, happen to have lower-level modules (or the ability to learn them). These agents will have a slightly higher fitness than the rest of the population, and will be selected as parents more often. When two such agents are paired together for reproduction, it is quite likely that one of the offspring will have two modules, one from each parent. These modules may also combine to form a higher-level module. In later generations, even larger modules can form, so that after a sufficient number of generations the low-level modules that were initially scattered randomly throughout the population have combined in single individuals. The genes of these individuals then begin to dominate the population due to an enhanced fitness, so that every individual has the high-level modules.

These simulations were repeated 100 times for each condition, varying the initial random seed. During the course of a simulation, the mean fitness of the population is monitored. Simulations were run for a maximum of 2000 generations, or until the population converged.

Results

In all three conditions, a sizeable proportion of the 100 populations converged on genotypes of maximum fitness. On average, trials in which agents were allowed either no learning or a moderate amount of learning outperformed those where less learning was allowed, as shown in Figure 3.

Discussion

The genetic operators of crossover are maximally suited to the hierarchical structure of the H-IFF problem. Unsurprisingly, crossover works well on this problem. Learning, which one might expect to perform as well or better, does not match the performance of the genetic operators alone. This result can be explained by considering the way that learning searches the space. Learning in the jester's cap is a mechanism for searching the neighborhood as determined by Hamming distance. This search is only effective for the lowest level of building blocks. At subsequent levels, local and global minima are close in recombination space, but not in Hamming space. At these higher levels, learning merely adds distractions to an otherwise successful algorithm, although adding a sufficient amount of learning can negate any detrimental effects.

In conclusion, learning in this type of hierarchical task is no more effective than genetic search because a way is needed to search module space, not Hamming space. As posed, the jester's cap assumes that low level modules are trivial to find. We next consider a sparse version of the task in which they are not so readily revealed.

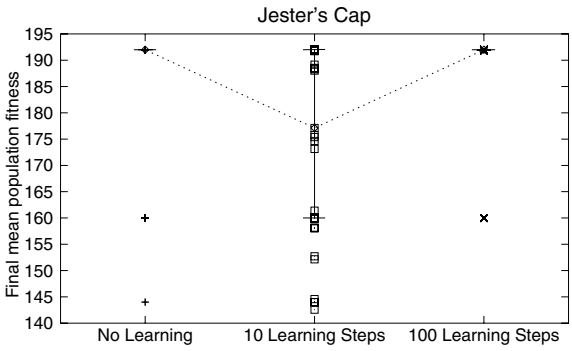


Figure 3: Performance of populations on the jester's cap task under three conditions. Each point represents the final mean fitness of a population. Error bars show first and third quartile and the medians are linked. Note that in the no learning (left) and moderate learning (right) conditions, the error bars are obscured because first, second (median) and third quartiles are all equal. Note also that many populations have identical mean fitness and tend to cluster around a set of discrete values because of the nature of the H-IFF function.

Simulation 2: The Sparse Jester's Cap

In the jester's cap simulations, rewards were given for modules of all levels (i.e, 1, 2, 4, 8, 16 and 32). In the sparse jester's cap, we consider rewarding only a subset of the levels. For example, in Figure 2 the blocks of size 2 may not contribute to the overall fitness of the solution. This modification allows us to vary the nature of the task from the maximally hierarchical H-IFF function (where all levels rewarded) to the NIAH function (where only the highest level rewarded). Varying the rewarded levels of the H-IFF function changes both the ease with which the initial modules can be found, as well as the ease with which the lower-level modules may be combined into the next higher-level of module. In the simulations in this section, only the building blocks of size 1, 16 and 32 are rewarded. With these choices of building-block and population sizes the smallest non-trivial modules (those of size 16) are difficult to find (*cf.* Hinton and Nowlan's simulations where the module is of size 20). It is thus expected that learning will substantially assist for the low-level modules. We repeated the first series of simulations using this alternative fitness function.

Results

Not surprisingly, the populations evolved using the sparse jester's cap fitness function fared substantially worse than those evolved with the (standard) jester's cap,

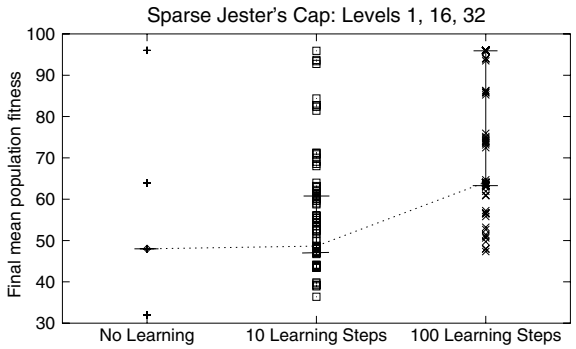


Figure 4: Performance of populations on the sparse jester's cap task under three learning conditions. Each point represents the final mean fitness of one population. Error bars show first and third quartile which are again obscured in some cases. Most populations with no learning (left) converged on a final fitness of 48, corresponding to one module. Populations in the moderate learning case converged on genomes giving a variety of fitness values, indicating some amount of residual learning in the genome (right). With a small amount of learning (center), performance was marginally improved over the no-learning case, although again residual learning remained.

as shown in Figure 4 (note that the y-axes differ between Figure 3 and Figure 4). In the condition of no learning, most populations (83 of 100) found a single module. With a small amount of learning, populations converged on marginally better solutions on average. In this condition, many residual question marks remained in the final populations. As a result, the individuals from these populations could only find modules with some degree of chance, causing the observed scattering of results in Figure 4. With moderate learning, some populations converged on the optimal genomes, others converged on genomes that gave agents the potential of finding the optimal solutions (i.e., there was again residual learning), while others converged on poor genomes. However, the populations with a moderate degree of learning, on average, outperformed the populations in the other learning conditions.

Discussion and Conclusions

This paper addressed the problem of how and when learning is an aid to evolutionary search in the jester's cap, a hierarchical modular task. Simulation 1 showed that evolutionary search could efficiently find the optimal solution with no learning. The addition of a small amount of learning detracted from this performance. A moderate amount of learning had no benefit, but did not detract either. It turns out that H-IFF is not a difficult task for a GA at the levels of complexity studied in this paper.

The main issues in reaching the highest levels of performance on H-IFF relate to maintaining a diversity of modules at intermediate levels. The population size of 500 in these simulations was clearly adequate for maintaining this diversity. Adding learnable alleles increases the search space, without a reciprocal benefit in assisting search.

Simulation 2 showed that the sparse jester's cap (1,16,32), a problem intermediate between H-IFF and NIAH, was not amenable to evolutionary search by the GA. The majority of populations only found a single module (fitness level 48). The smallest module involved 16 bits, and the likelihood of finding two such modules in any one population before convergence was minimal.

In contrast to Simulation 1, in Simulation 2 a small amount of learning (ten steps) marginally improved the success with which populations discovered modules. Ten learning steps is sufficient to effectively search four to five learnable genes, and in this case, learning clearly did provide a reciprocal benefit that more than compensated for the increase in the search space.

Increasing the learning from ten to 100 steps substantially improved the success of populations. All populations found at least one module, 75% found two modules, and at least 25% reached optimal performance. One hundred learning steps is sufficient to search six to seven learnable genes. Although this is only two more than searched by ten learning steps, it had a demonstrable effect on performance.

Simulation 2 demonstrated that in the sparse version of the jester's cap, learning is required to discover the modules, as in NIAH. As in Hinton and Nowlan's simulations, populations are unlikely to find high-level modules by crossover alone. Learning is able to guide the population towards finding the low level modules, and then crossover combines them. However, the performance is still not optimal, and room for improvement remains.

In conclusion, there appears to be a role for learning in situations where crossover is an ineffective search technique. Crossover searches module space whereas learning searches Hamming space. In tasks such as the jester's cap there is very little need for searching Hamming space and the majority of optimization can be effectively performed in module space. In this task, Hamming search is useful only at the lowest-level of module. For higher-level modules crossover searches through combinations of peaks, rather than traversing the troughs between them (Figure 1). Local search provides the wrong operator for preserving and improving fitness because it spends too much time in the troughs of fitness space. In the sparse jester's cap, modules must be discovered before searching module space becomes a viable approach. The difficulty in finding these modules necessitates local search.

NIAH and H-IFF may be viewed as being on alternative ends of a spectrum. In the former, the (single) module is difficult to find as there is no partial feedback to guide search. Learning is required to act as a proxy for this partial feedback. In the latter, modules abound so the important factor is not finding the modules, but discovering how to put them together. Consequently, the

best search process is one that searches through combinations of modules rather than searching for the modules themselves. In this case, learning is merely a hindrance. The sparse jester's cap represents an intermediate point on the spectrum. Modules are sufficiently difficult to find so that learning is required to give partial feedback in the search for the lowest-level modules. Once the modules have been found, recombination can function.

Acknowledgements

We thank Rik Belew and two anonymous reviewers for their constructive feedback on this paper. This research has been supported by a CSEE scholarship to JRW.

References

Baldwin, J. M. (1896). A new factor in evolution. *American Naturalist*, 30:441–451. Reproduced in Belew, R. K. & Mitchell, M. (Eds.), *Adaptive Individuals in Evolving Populations*. Addison-Wesley, Reading, MA.

Belew, R. K. (1990). Evolution, learning and culture: computational metaphors for adaptive search. *Complex Systems*, 4(1):11–49.

Dawkins, R. (1986). *The Blind Watchmaker*. Penguin Books.

Forrest, S. and Mitchell, M. (1993). Relative building-block fitness and the building-block hypothesis. In Whitley, D., editor, *Foundations of Genetic Algorithms*, volume 2, pages 109–126. San Mateo, CA: Morgan Kaufmann.

French, R. and Messinger, A. (1994). Genes, phenes and the baldwin effect: Learning and evolution in a simulated population. In Brooks, R. A. and Maes, P., editors, *Artificial Life IV*, pages 277–282.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, MA.

Harvey, I. (1993). The puzzle of the persistent question marks: A case study of genetic drift. In Forrest, S., editor, *Proceedings of the Fifth International Conference on Genetic Algorithms*, pages 15–22, San Mateo, CA. Morgan Kaufmann.

Hinton, G. and Nowlan, S. (1987). How learning can guide evolution. *Complex Systems*, 1:495–502.

Holland, J. H. (1992). *Adaption in Natural and Artificial Systems*. MIT Press, 2nd edition.

Holland, J. H. (2000). Building blocks, cohort genetic algorithms, and hyperplane-defined functions. *Evolutionary Computation*, 8(4):373–391.

Mayley, G. (1996). The evolutionary cost of learning. In Maes, P., Mataric, M. J., Meyer, J.-A., Pollack, J., and Wilson, S. W., editors, *Proceedings of the Fourth International Conference on Simulation of Adaptive behavior: From Animals to Animats 4*. MIT Press/Bradford Book.

Mitchell, M. (1996). *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA.

Mitchell, M. and Belew, R. K. (1996). Preface to 'How learning guides evolution' by G. E. Hinton & S. J. Nowlan. In *Adaptive Individuals in Evolving Populations: Models and Algorithms*, volume XXVI of *Santa Fe Institute Studies in the Science of Complexity*, pages 443–446. Addison-Wesley.

Mitchell, M., Forrest, S., and Holland, J. H. (1992). The royal road for genetic algorithms: Fitness landscapes and GA performance. *Proceedings of the First European Conference on Artificial Life*, pages 245–254.

Mitchell, M., Holland, J. H., and Forrest, S. (1994). When will a genetic algorithm outperform hill climbing. In Cowan, J. D., Tesauro, G., and Alspector, J., editors, *Advances in Neural Information Processing Systems*, volume 6, pages 51–58, San Mateo, CA. Morgan Kaufmann Publishers, Inc.

Morgan, L. C. (1896). On modification and variation. *Science*, 4:733–740.

Pelikan, M. and Goldberg, D. E. (2000). Hierarchical problem solving by the bayesian optimization algorithm. IlliGAL Report No. 2000002, Illinois Genetic Algorithms Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL.

Pereira, F. B., Machado, P., Costa, E., Cardoso, A., Ochoa-Rodriguez, A., Santana, R., and Soto, M. (2000). Too busy to learn. In *Proc. of the 2000 Congress on Evolutionary Computation*, pages 720–727, Piscataway, NJ. IEEE Service Center.

Tooby, J. and Cosmides, L. (1994). Origins of domain specificity: The evolution of functional organisation. In Hirschfeld, L. and Gelman, S., editors, *Mapping the Mind*, pages 85–116. Cambridge University Press.

Watson, R., Hornby, G., and Pollack, J. (1998). Modeling building-block interdependency. *Parallel Problem Solving from Nature, proceedings of the Fifth International Conference*, pages 97–106.

Watson, R. and Pollack, J. (1999). Hierarchically-consistent test problems for genetic algorithms. In Angeline, P. J., Michalewicz, Z., Schoenauer, M., Yao, X., and Zalzala, A., editors, *Proceedings of 1999 Congress on Evolutionary Computation*, pages 1406–1413. IEEE Press.

Wiles, J., Schulz, R., Bolland, S., Tonkes, B., and Hallinan, J. Selection procedures for module discovery: Exploring evolutionary algorithms for cognitive science. This volume.

Supporting Understanding through Task and Browser Design

Jennifer Wiley (jwiley@uic.edu)
Department of Psychology, 1007 W. Harrison Street (M/C 285)
Chicago, IL 60607 USA

Abstract

While electronic text offers the potential to explain, illustrate, and scaffold understanding in powerful new ways, few studies on educational use of electronic text resources have shown significant learning gains, or even measured learning outcomes in controlled experiments (Chen & Rada, 1996; Dillon & Gabbard, 1998). In a follow-up to previous studies (Wiley & Voss, 1999), the present experiments study the effects of different tasks and browser designs on navigation and reading patterns, as well as on memory and comprehension measures from electronic text. These studies have revealed that only when both the task and the design support integration (such as in a two-windowed browser) and students are explicitly directed how to use the feature, do students take advantage of the flexibility of the multiple-source environment, integrate across sources, and achieve the best level of understanding.

Introduction

One promise of using electronic text in the classroom is the potential for students to search for, access and read multiple forms of information about a topic. Since the search for and navigation of digital documents is student-initiated, requires student interaction, captures the student's interest, proceeds at the student's own pace, and allows for flexible navigation and juxtaposition of multiple sources, a number of theorists have suggested that the web might be a powerful tool for student instruction (Beeman, et al 1987, Spiro & Jehng, 1990). This optimism is consistent with a number of recent cognitive studies demonstrating that activities that require readers to engage in active, constructive and integrative tasks lead to the best understanding of the subject matter (e.g., Chi, de Leeuw, Chiu & LaVancher, 1994; McNamara, et al., 1996).

However, a review of the literature on educational use of electronic text yields two striking conclusions. First, students generally fail to utilize hypertext links and multiple window capabilities effectively, if at all, as they read (Foss, 1989; Gordon, et al., 1988). This is especially true of novice users (Foltz, 1996; Gray & Shasha, 1989; Tombaugh, Lickorish & White, 1987). And second, few studies on educational use of electronic documents, whether from stand-alone hypermedia or the World Wide Web, have actually shown significant learning gains (Chen & Rada, 1996; Dillon & Gabbard, 1998). What is critically needed is for experiments to determine which specific instructional contexts may allow for effective educational use of electronic text.

Although there has been a great deal of evaluation on effective browser design from a Human-Computer Interaction (HCI) standpoint, effectiveness has been measured largely in terms of efficiency of search or ease in information finding (Dillon & Gabbard, 1998). While such fluency measures may be related to some extent to the amount of information a person is able to recall after reading from computer screens, they may not be correlated with whether a person develops an understanding of the text that is being read. A number of studies have shown that conditions that produce the best surface memory for text are not the best conditions for producing the best understanding of text (e.g. McNamara, et al., 1996; Mayer, 1999; Wiley & Voss, 1999). While surface memory may correlate with the fluency or ease of information processing, understanding may depend to some extent on the need to put effort into developing an underlying representation or situation model of the text (Kintsch, 1998). Thus, previous assessments from an HCI perspective cannot reliably indicate which screen layouts will be most effective for promoting understanding from electronic text. Given this educational goal, browser design must be evaluated specifically using measures of conceptual learning.

In a review of the published studies on hypermedia and learning outcomes between 1990 and 1996, Dillon and Gabbard found only 11 studies that performed controlled experiments on hypermedia and learner performance. Of those 11 studies, there were only four results that actually seemed to suggest an advantage for learning from hypermedia over paper. In the majority of studies there was no clear difference between learning from hypermedia and learning in a control (more traditional) setting. While this may be viewed more optimistically as evidence that learning from electronic resources may sometimes be no worse than traditional classroom methods, there is hardly an overwhelming body of evidence that the web can generally be relied upon to provide an enriching instructional experience.

Of the four studies that netted positive learner performance, only two seem to indicate that hypertext may allow students to engage in learning at a more conceptual level. One of these reported that students learned to recognize aircraft more efficiently and effectively when they were able to view the images

next to each other (and even overlay two images) in a browser during learning (Psotka, Kerst & Westerman, 1993). Clearly the ability to juxtapose and overlay images gave subjects a better concept of the aircraft prototypes, and this allowed them to perform well on later recognition tasks. In a second study, Jacobson and Spiro (1995) found that a hypermedia environment as opposed to a linear electronic presentation of the same materials allowed for the best performance on a problem-solving essay task. Interestingly, the linear presentation of the same material allowed for the best recall of the facts. Unfortunately, no converging evidence of better comprehension in the hypertext vs. linear condition was obtained. However, the differences in essay quality suggest that students in the hypertext conditions benefitted from the flexibility and ability to jump between sources in the hypertext format, allowing for better synthesis of the material that was presented. This result is consistent with perhaps the most consistently cited study on learning in hypertext (Egan et al, 1989) which found that a multi-window environment called Superbook led to better essay writing than a paper control condition.

Based on their review, and these last two studies in particular, Dillon and Gabbard suggest that hypermedia may afford particular advantages for learners on tasks that require comparison across sources. Notably, when students are not given the ability to view two sources at once in a computer environment, they perform less well than with paper on tasks which require integration across two documents (Wang & Liebscher, 1988). Consistent with the intuition provided by these previous studies, Wiley & Voss (1999) reported that students can show better conceptual learning from a web-like environment when they are provided with multiple windows and are given a task that requires them to integrate information across sources.

In one of few empirical studies evaluating conceptual learning from a web-like environment, Wiley and Voss (1999) demonstrated that reading multiple sources presented in a browser can lead to better understanding of subject matter than reading from a textbook. In this study, students read about the Irish Potato Famine either from several on-line documents in a two-window web site or they read the same information in the form of a textbook chapter. When students were asked to write an argument of "What produced the significant changes in Ireland's population" instead of a narrative, and read the on-line sources instead of the textbook chapter, students gained the best understanding of the material. Understanding was assessed by the causal and integrated nature of their essays, as well as their performance on inference verification and analogy identification tasks. Wiley and Voss (1999) concluded that tasks which require students to construct their own representations of a situation will yield the most conceptual learning in web-like environments. The argument writing task promoted understanding because it required students to integrate information from across the multiple sources as they created support for a thesis. And, the multiple source condition may have promoted understanding by supporting the comparison and integration of the individual sources.

This is an important finding, demonstrating empirically that electronic text can be an effective tool for developing student understanding. However, it is important to note the very specific circumstances under which better understanding may be achieved from web-like environments. For one, only students given a task requiring integration of information across sources showed better learning from a browser. Otherwise, learning from a browser was actually poorer than from a textbook. In fact, there were a number of particular features of the Wiley & Voss (1999) environment, any of which may be important in order for the effect to be obtained:

The site had a small set of documents.
The documents were selected for the user.
The documents were largely relevant.
There were no links embedded in the texts.
Each document fit in a single window.
The task was well-defined and specific.
The task required integration across sources.
The browser had two side-by-side windows.
The overview menu was accessible through an icon.
Images were presented in their own window.
Students were instructed to use both windows.
Students were instructed how to use the menu icon.
Conceptual learning was assessed in the post-test.

The present experiments directly test whether the design of the browser with two side-by-side windows might have been critical for the better learning in the web source/argument writing task condition. Although many computerized tutors and interactive environments use multiple windows, there has been little work on how students use multiple windows or the optimal conditions for multiple window use (von Oostendorp, 1996). Interestingly, the three other studies that suggest that students can gain better conceptual understanding from hypermedia environments (Egan, et al., 1989; Pstoka, et al 1993, Jacobson & Spiro, 1993) all used a multi-window display. However, there have been no studies that have manipulated the number of windows and directly measured comprehension. With converging data from essay tasks, comprehension tests and eyetracking protocols, the present studies address whether a multi-window browser supports better understanding in a web-like environment.

Experiment 1

The first experiment investigated the effect of two-window browsers on learning historical subject matter from a web site. This experiment tested the hypothesis that the design of the browser had an impact on students' understanding. Thirty undergraduates were asked to read 10 pages from a web site about the Irish Potato Famine in order to write either a narrative or an argument of what produced the significant changes in Ireland's population. (The pages contained 5 texts (about 1500 words), 4 tables, 1 graph and 1 map.) In addition, students either read the information from a single-window browser, a two-window browser, or from a two-window browser with specific instructions about why they were being given two windows. Further in this third condition the list of documents was split across the two windows, so that in order to read all of the information readers had to use both windows.

Student learning was assessed with a number of learning measures taken from Wiley and Voss (1999). Of most interest are three measures thought to reflect understanding: the proportion of sentences in student essays which represented an integration or "transformation" of the presented information (as opposed to simply copying the presented information), an inference verification task and an analogy task. The inference verification task (IVT) contained 10 inferences that could be generated by integrating information across two sources, such as "As rent costs increased, emigration from Ireland increased." as well as 10 distractors. The analogy task consisted of short descriptions of potentially analogous events. Students were asked to rate the similarity of the causes of each event with the population decline in Ireland. These analogies were intended to vary in surface and deep similarity, and the critical analogy, the institution of a Poll Tax in the U.S. South after the Civil War, was intended to be similar only on a deep level (as it was related to sociopolitical inequities and class power struggles, but there was no large-scale loss of life). Thus, recognition of the Poll Tax as causally similar to the changes in Ireland's population indicates a particularly good understanding of the text. This rating serves as the critical analogy task (CAT).

The means for each condition on each measure are presented in Figure 1. Performance on all measures was better when students wrote arguments as opposed to narratives (TRSENT: $F(1,24)=14.9$, $p<.01$; IVT: $F(1,24)=10.9$, $p<.01$; CAT: $F(1,24)=11.23$, $p<.01$). The main effect for number of windows only reached significance for the inference task, $F(2,24)=5.3$, $p<.01$, as did the interaction, $F(2,24)=7.89$, $p<.01$. However, the interaction approached significance for the proportion of transformed sentences ($p<.15$) as did the main effect for number of windows on the critical analogy task ($p<.13$). Pairwise comparisons based on a priori hypotheses revealed that the two-window/ argument writing condition significantly outperformed

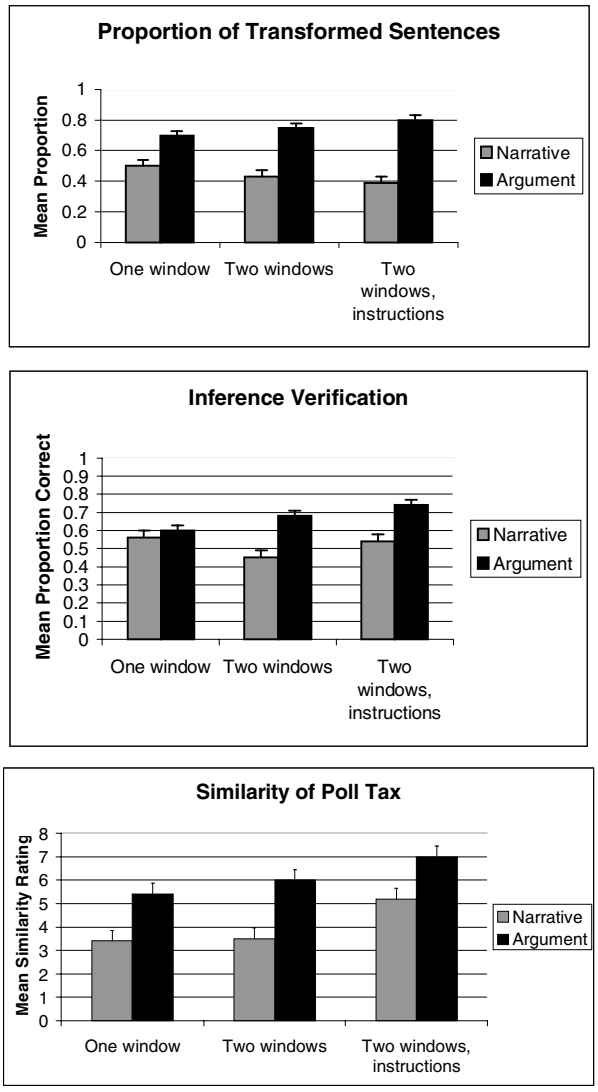


Figure 1: Experiment 1 Results.

the single window/narrative writing conditions on all measures ($ps<.05$), but only when students were given specific instructions.

Two other sources of data may be of interest here. The first is anecdotal evidence from a failed pilot study showing that students are quite reluctant to use a two-window browser. In the first design of the two-window browser, I failed to set the frame so that it could not be removed. The first thing that students did when they were seated in front of a computer with the two-window browser was click and drag that frame off the screen so they had one big window. It is important to note that I designed all my materials so that they would fit in a single side-by-side window. Thus, it was not that students could not see all of the documents that prompted them to want a bigger screen. Their initial preference was for a single window. Further, even once the two-window browser was set so students couldn't remove the frame, many of them still failed to ever use the second window. Instead, many students kept the graph in the right window and read through the sources

on the left. Only when students were given the specific instructions and forced to use both windows did all of the students actually use both windows.

Previous work with multiple window environments is consistent with these findings. Foss (1989) found that about a third of people generally prefer simple screens and only having a single window open. This prevents students from making important connections across sources. Another third of readers did open more than one window at once, but they did not use the multiple windows effectively. Their screens quickly became messy and busy. Only a third of Foss' subjects were able to use multiple windows effectively for a search and decision making task. Tombaugh et al (1987), also found a general preference for and better search in a single window environment, and found that only with practice in a multiple window environment were participants able to use the overlapping multiple windows efficiently. Part of these results may be due to the overlapping window environment that was used, however, as Instone, et al (1996) found that participants were able to use a tiled multiple-window environment more efficiently than an overlapping window display. Unfortunately, all of these previous experiments measured the effects of the windows in terms of speed and accuracy of search for information, and not in terms of comprehension. Only the present study has manipulated the number of windows and investigated how the use of multiple windows can lead to a better conceptual understanding of the subject matter, as it allows for the concurrent presentation of related concepts from different sources.

A second additional source of information about how readers behaved in the different conditions comes from analysis of browsing logs and eyetracking data. These sources indicated that both general instruction about how to use web sources and specific writing instruction yielded different navigation and reading patterns. A pilot eyetracking study on 4 students in the 2-window/2-list condition gives us a better idea of exactly how the students used these windows. Two of these students were told to read with the purpose of writing a narrative, and the other two were told the argument instruction. Like all other students, these readers tended to begin their task by reading through each document, one at a time, in order. Students usually simply went down the documents in the list on the left side of the screen and then down the documents on the right side of the screen. During this initial reading phase, the eyes rarely left the document that was being read. At this point, both readers in the narrative condition stopped reading and declared they were ready to write their essay. On the other hand, both students in the argument condition moved on to a second phase of reading.

One student started over again from the beginning, starting at the top of the left window list and skimmed the documents in order. But, from time to time, she would call up documents in the other window, or skip to another document on the same list and look for particular sentences. The other student in the argument condition began the second phase by calling up pairs of documents and alternated reading between the two. Importantly, the sentences that these students tended to re-read were important for inferences about the causes of the Potato Famine. Eyetracking data revealed that students in argument condition spent more total time on sentences important for inferences. Thus, the selection patterns and eyetracking evidence suggest that under some conditions web sites can promote more active reading patterns, suggesting more active integration of the text at a conceptual level. Further, this second phase of reading, or review of documents, seems to be particularly critical for understanding. It is what students do when they are reading from paper documents, and what Dee-Lucas and Larkin (1995) found when students effectively used structured overviews in electronic text.

Taken together, these sources of data demonstrate that students need to have both a task and an environment that forces them to be more active in order for students to gain the benefits of web resources. Only when both the task and the design support integration, and students are explicitly directed how to use the feature, do students take advantage of the flexibility of the multiple source environment. Only then do students integrate across sources, selectively re-read sources, and achieve the best level of understanding.

The second experiment is an important extension of this work using scientific texts as content. Of particular interest is whether the ability to juxtapose two documents, while performing a task that supports integration, will allow for better understanding of scientific concepts as well.

Experiment 2

Although there may be some differences between reading from history and science-related text, when readers must connect information across documents, in order to make inferences or construct global models of causality, then simultaneous presentation of the sources that need to be linked should help regardless of the subject matter. Thus, the second experiment investigated the effect of two-window browsers on learning from a scientific web site. This experiment tested the hypothesis that the design of the browser has an impact on student's scientific understanding. Forty undergraduates were asked to read 16 pages from a web site about Earthquakes and Volcanoes (based on sources from the USGS web site) in order to write either an essay or an argument of "What caused the explosion of Mt. St. Helens?" (The pages contained 10 texts (about 3000 words), 4 diagrams, 3 maps, and 2 photographs.) In addition, students either read multiple sources in a single-window browser, or multiple sources in a two-window browser with specific

instructions about why they were being given two windows. Further in the two-window condition the list of documents was split across both windows, so that in order to read all of the information readers had to use both windows. This yielded a 2x2 (writing task x presentation format) design with 10 students in each of the conditions.

Student learning was assessed with a number of learning measures similar to those used in Wiley & Voss (1999) and in Experiment 1. The same 3 measures of understanding are reported as in Experiment 1: proportion of transformed sentences (TRSENT), the inference verification and analogy tasks. The inference verification task (IVT) contained 10 inferences that could be generated by integrating information across two sources, such as "Volcanoes are likely to develop where continents collide" as well as 10 distractors. The critical analogy task (CAT) asked students to rate the similarity of the causes of the Kobe earthquake with the causes of the Mt. St. Helens eruption. The Kobe earthquake was intended to be similar only on a deep level (as it was related to disturbance due to subduction of a tectonic plate, but there was no volcanic activity). Thus, recognition of Kobe as causally similar to Mt. St. Helens indicates a particularly good understanding of the text, and the relation between plate tectonics and volcanic activity.

The means for each condition are presented in Figure 2 for the 3 tasks. As in Experiment 1, performance on all tasks was better when students wrote arguments (TRSENT: $F(1,36)=8.96$, $p<.01$; IVT: $F(1,36)=9.06$, $p<..01$; CAT: $F(1,36)=8.81$, $p<.01$). The main effect for number of windows only reached significance for the proportion of transformed sentences, $F(1,36)=6.82$, $p<.01$.

While none of the interactions were significant, based on previous results and a priori hypotheses, pairwise comparisons were performed and revealed that the two-window/argument condition outperformed the essay conditions on all measures ($ps<.08$). However, it is notable that although there was a trend toward a main effect for number of windows on the inference task ($p<.14$), there were no trends in either the windows effect or the interaction on the critical analogy task. This suggests that while the best essay writing may have occurred due to combination of writing task and two-window browser, for the learning measures, the writing task alone was responsible for better understanding.

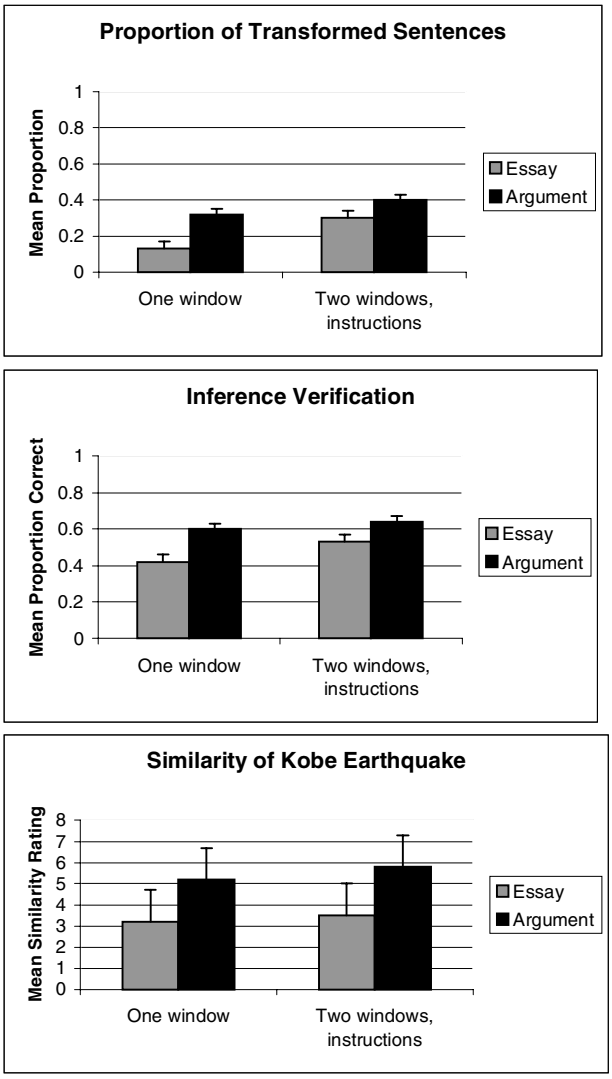


Figure 2: Experiment 2 Results.

Conclusions and Implications

In both experiments, when both the task and the browser design supported integration, and there was explicit instruction how to use the two-window browser, students were able to write more integrated essays in a multiple window environment. This condition also led to the best conceptual learning in both experiments. Although in Experiment 1 it seemed both the writing task and the number of windows contributed to the effect, in Experiment 2, the effect of the two-window browser on the understanding of the subject matter was not as evident. The argument writing task, however, did support better understanding in both history and science.

There are many possible reasons for the differential effects of the browser design in these two studies. The most obvious are the differences in the materials that students were presented with. More information was presented in the scientific site. Further, looking at the single window/narrative essay condition as a baseline, we can see that the scientific tests were more difficult for students. Students accurately recognized 55% of inferences in the history test, whereas they recognized only 40% of inferences correctly in the science site. Further, different kinds of images were used in the history and science sites. These findings have led to several interesting questions that we are currently

pursuing (Wiley, Ash, Brodhead & Sanchez, 2001). Do students simply respond to history and science subject matter differently? Are images processed differently in the two domains? Did the different types of images in the two sites lead to learning differences? Or, is the difficulty of the subject matter driving these processing differences?

Even though we are still looking for the best task/environment combination for conceptual learning from scientific web sites, taken together, the present studies demonstrate that specific conditions are necessary for effective educational use of electronic texts. In order for conceptual learning to occur, readers of electronic text may need a multimedia environment that promotes integration of the presented information and certainly need a task that prompts integration across sources. Only through the specification and demonstration of which computerized learning environments lead to better understanding, may we begin to realize some of the educational potential of electronic text.

Acknowledgments

This work was supported by grants from the Office of Naval Research and the Paul G. Allen Virtual Education Foundation. Thanks to Linda Gentry, Ed Kearney, and Rebecca Schrader for their research assistance.

References

Beeman, W., Anderson, K., Bader, G., Larkin, J. McClard, A., McQuillian, M. & Sheilds, M. (1987) Hypertext and pluralism: from lineal to nonlineal thinking. In *Proceedings of Hypertext 87*, pp. 67-88, University of North Carolina, Chapel Hill.

Chen, C. & Rada, R. (1996) Interacting with hypertext: A meta-analysis of experimental studies. *Human Computer Interaction, 11*, 125-156.

Chi, M., de Leeuw, N., Chiu, M. & LaVancher, C., (1994). Eliciting self-explanations: improves learning. *Cognitive Science, 18*, 439-478.

Dee-Lucas, D. & Larkin, J..(1995). Learning from electronic texts: Effects of overviews for information access. *Cognition & Instruction, 13*, 431-468.

Dillon, A. & Gabbard, R. (1998). Hypermedia as educational technology: A review of the quantitative research literature on learner comprehension, control and style. *Review of Educational Research, 68*, 322-349.

Egan, D., Remde, J., Landauer, T., Lochbaum, C. & Gomez, L. (1989) Behavioral evaluation and analysis of a hypertext browser. *Proceedings of CHI 89*, 205-210.

Foss, C. (1989). Detecting lost users: Empirical studies on browsing hypertext. INRIA Tech Report 973. Valbonne, France: L'Institut National de Recherche en Informatique et en Automatique.

Foltz, P. (1996). Comprehension, coherence and strategies in text and hypertext. In Rouet, J.F. et. al. (Eds) *Hypertext and Cognition.* Mahwah, NJ: Erlbaum.

Gordon, S., Gustavel, J., Moore, J. & Hankey, J. (1988). The effects of hypertext on reader knowledge representation. *Proceedings of the Human Factors Society 32nd Annual Meeting*, 296-300.

Gray, S. & Shasha, D. (1989) To link or not to link *Behavior Research, Methods, Instruments, and Computers, 21*, 326-333.

Instone, K., Teasley, B. & Leventhal, L. (1996) Lessons learned from redesigning hypertext user interfaces. In van Oostendorp & de Mul (Eds*) Cognitive aspects of electronic text processing*. Norwood, NJ: Ablex.

Jacobson, M. & Spiro, R. (1995) Hypertext learning environments, cognitive flexibility and the transfer of complex knowledge. *Journal of Educational Computing Research*, *12*, 301-303.

Kintsch, W. (1998*) Comprehension: A paradigm for cognition*. Cambridge: Cambridge.

Lehto, M., Zhu, W. & Carpenter, B. (1995) The relative effectiveness of hypertext and text. *International Journal of Human Computer Interaction, 7*, 293-313.

McNamara, D. Kintsch, E., Songer, N., & Kintsch, W. (1996) Are good texts always better*? Cognition & Instruction, 14*, 1-43.

Psotka, J., Kerst, S. & Westerman, T. (1993) The use of hypertext and sensory-level supports in visual learning. *Behavior Research Methods, 25*, 168-172.

Spiro, R. & Jehng, J. (1990) Cognitive flexibility and hypertext. In D. Nix & R. Spiro (Eds*) Cognition, Education and Multimedia.* Hillsdale, NJ: Erlbaum.

Tombaugh, J., Lickorish, A. & Wright, P. (1987) Multi-window displays for readers of lengthy texts. *International Journal of Man-Machine Studies, 26*, 597-615.

van Oostendorp, H. (1996) Studying and annotating electronic text. In J.F. Rouet, et al (Eds*.) Hypertext and Cognition*. Mahwah, NJ: Erlbaum.

Wang, X. & Liebscher, P. (1988) Information seeking in hypertext: Effects of physical format and search strategy. *Proceedings of the ASIS Annual Meeting, 25*, 20-205.

Wiley, J., Ash, I., Brodhead, A. & Sanchez, C. (2001, May) *The impact of images in learning from web pages on science and history*. Paper presented at the Midwestern Psychological Association, Chicago, IL.

Wiley, J. & Voss, J. F. (1999) Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology, 91*, 1-11.

Access to Relational Knowledge: a Comparison of Two Models

William H. Wilson (billw@cse.unsw.edu.au)
Nadine Marcus (nadinem@cse.unsw.edu.au)
School of Computer Science and Engineering, University of New South Wales, Sydney, New South Wales, 2052, Australia

Graeme S. Halford (gsh@psy.uq.edu.au)
School of Psychology, University of Queensland, Brisbane, Queensland, 4072, Australia

Abstract

If a person knows that Fred ate a pizza, then they can answer the following questions: Who ate a pizza?, What did Fred eat?, What did Fred do to the pizza? and even Who ate what? This and related properties we are terming *accessibility properties* for the relational fact that Fred ate a pizza. Accessibility in this sense is a significant property of human cognitive performance. Among neural network models, those employing tensor product networks have this accessibility property. While feedforward networks trained by error backpropagation have been widely studied, we have found no attempt to use them to model accessibility using backpropagation trained networks. This paper discusses an architecture for a backprop net that promises to provide some degree of accessibility. However, while limited forms of accessibility are achievable, the nature of the representation and the nature of backprop learning both entail limitations that prevent full accessibility. Studies of the degradation of accessibility with different sets of training data lead us to a rough metric for learning complexity of such data sets.

Introduction

The purpose of this research is to determine whether a backpropagation net can be developed that processes propositions with the flexibility that is characteristic of certain classes of symbolic neural net models. This has arguably been difficult for backpropagation nets in the past. For example, the model of Rumelhart and Todd (1993) represents propositions such as "canary can fly". Given the input "canary, can" it produces the output "fly". However processing is restricted, so it cannot answer the question "what can fly?" ("canary").

There are, however, at least two types of symbolic nets that readily meet this requirement. One type of net model makes roles and fillers oscillate in synchrony (Hummel & Holyoak, 1997; Shastri & Ajjanagadde, 1993) while another is based on operations such as circular convolution (Plate, 2000) or tensor products (Halford, et al., 1994; 1998; Smolensky, 1990). These models appear to have greater flexibility than models based on backpropagation nets, in that they can be queried for any component of a proposition. We will refer to this property of tensor product nets as omni-directional access (cf. Halford, Wilson & Phillips, 1998). Omni-directional access is the ideal form of accessibility.

Another reason for investigating this lies in the work of Halford, Wilson, and Phillips (e.g. 1998) which seeks in part to define a hierarchy of cognitive processes or systems and to draw parallels between this hierarchy and a second hierarchy of types of artificial neural networks. Levels 0 and 1 of this second hierarchy are 2- and 3-layer feedforward nets, and levels 2-5 are tensor product nets of increasing rank. It thus becomes interesting to consider how well feedforward nets can emulate tensor product networks.

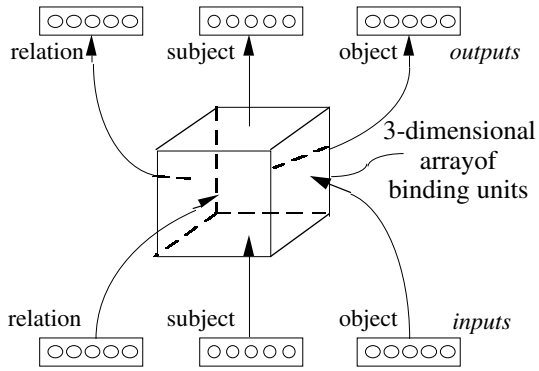


Figure 1 – Tensor product network of rank 3.

As tensor product networks are not as well known as feedforward networks, we shall describe them and their accessibility properties briefly here before proceeding. Tensor product networks are described in more detail, and from our point of view, in Halford et al. (1994). Briefly, a rank k tensor product network consists of a k-dimensional array of "binding units", together with k input/output vectors. For example, a rank 2 tensor product network is a matrix, plus 2 input/output vectors. To teach the network to remember a fact (that is, a k-tuple), the input/output vectors are set to be vectors representing the components of the k-tuple, and a computation is performed that alters the k-dimensional array. Subsequently that fact can be accessed in a variety of ways. It is common to interpret the first

component of the *k*-tuple as a predicate symbol, and the remaining components as argument symbols - e.g. for rank 3, the components might be vectors representing the concepts *likes jane pizza* (Jane likes pizza) (see Figure 1). Once this fact has been taught to a rank 3 tensor product network, the following 7 queries can be formulated and answered by a computation involving the tensor product network.

1) Is *likes(jane,pizza)* true?
2) Who likes pizza? This we often write as *likes(X, pizza)*? The response depends on what else has been taught to the tensor product network. If the tensor product network also knows that *likes(fred,pizza)* and *likes(mary,pizza)* then the response will be the sum of the vectors representing Jane, Fred, and Mary - often written jane + fred + mary.
3) What does Jane like? - *likes(jane,X)*? Similar to 2).
4) What relationships hold between Jane and pizza? - *X(jane,pizza)*? Again, similar to 2).

These four are referred to as limited accessibility.

5) Who likes what? - *likes(X,Y)*? The response in this case would be a rank 2 tensor product network storing the pairs (X,Y) for which *likes(X,Y)* is known to the original rank 3 tensor product network. The tensor product network approach solves this by producing a rank 2 tensor product network, which stores the pairs (X,Y). (This output possibility, and corresponding ones for 6) and 7) below, are not shown in Figure 1).
6) Who does what to pizza? - *X(Y,pizza)*? Like 5).
7) Jane does what to what? - *X(jane,Y)*? Like 5).

The full set of 7 forms of access are referred to as full accessibility, or omni-directional access.

A rank 4 tensor product network would have 15 access modes, a rank 5 tensor product network would have 31 access modes, and so on. Provided that an orthonormal set of vectors is used for the set of vectors representing concepts, retrieval is perfect. Facts are learned by a tensor product network one at a time, and do not interfere with each other (given orthonormal representation vectors).

Tensor product networks using orthonormal sets of representation vectors exhibit what has been called full omni-directional access to the facts that have been taught, as noted above. Humans attempting similar tasks may find some types of access easier than others. For example, children who have recently learned sets of multiplication facts such as 9×7=63 are able to use this knowledge to perform division (9×X= 63, what is X?), but may find this more difficult than multiplication (9×7=X, what is X?). We use the term *accessibility* to refer to imperfect or partial versions of omni-directional access. It turns out that some of the nets discussed in this paper also exhibit accessibility rather than full omni-directional access.

Our specific aim in this paper is to experiment with a feedforward net design that appears to have potential to provide at least limited accessibility in a rank 3 situation. When we move to feedforward networks trained by error backpropagation, we hope to preserve the accessibility property that is characteristic of symbolic nets. The model resembles an auto-encoder but has restricted connectivity.

Architecture of the network

The particular backpropagation network we used to test for accessibility consisted of the following components: 15 input units, 15 hidden units and 15 output units. The 15 input units were used to represent 3 items or patterns, each made up of 5 elements. The hidden and output units also each consisted of three groups of 5 units, connected as shown in Figure 2. The input patterns represented relational instances of the form RELATION(SUBJECT, OBJECT). The target output contained the same information: namely, RELATION, SUBJECT and OBJECT.

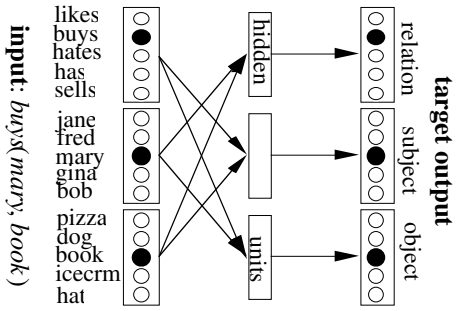


Figure 2 - Connections in our feedforward net architecture.

Notice that this network consists of three functions: one takes as inputs a relation name and a subject, and produces an object as output, the second takes relation name and object and produces subject as output, and the third takes subject and object and produces relation name as output. Thus, while it resembles a traditional auto-association net, note that regular auto-association nets allow connection paths between corresponding input and output neurons, typically allowing total interconnection between input and hidden layers, and between output and hidden layers. In essence, the network architecture can be unraveled into 3 distinct networks that share common inputs. Thus, any weight in the network is influenced by the output errors in only one of the 3 output sets (relation, subject, object). The net makes learning easier by constraining the learning algorithm to look for sets of weights that, for example, ignore predicate input when trying to infer predicate output from argument input. We also conducted some pilot studies with a fully connected network and the network's performance was inferior.

Notice that both the tensor product net architecture and the architecture we are studying here have three groups of input and three groups of output nodes.

Experimental design

The network was given three different sets of relational instances to learn. Each set was made up of five different relational instances. For example, given the relational instance *likes(jane,pizza*), *likes* is the RELATION, *jane* is the SUBJECT and *pizza* is the OBJECT. The training sets varied in terms of the amount of overlap or the degree of interaction between the elements of each of the five relational instances. Training set 1 was set up to contain little or no overlap between the different relational instances. Training set 3 consisted of five relational instances with a large degree of interaction between the different instances. Training set 2 contained an intermediate degree of overlap.

It was hypothesized that relational instances with the least amount of interaction between the different instances would be the easiest to learn. This is because each instance does not have components that overlap with the other instances. These relations are one-to-one mappings. Accordingly, the network is most likely to achieve success in learning such a set of facts. In contrast, the set of relations with the highest level of overlap is expected to be the most difficult for the network to learn. These relations can be classified as many-to-many mappings, and so cannot be completely learned by a feedforward network. Accessibility may be easier to obtain if each relational instance can be represented in isolation, with little or no reference to the other relations. As the overlap and interaction between the relations and their elements increases, so the degree of accessibility that can be obtained is likely to decrease. This is because information from other related instances is more likely to interfere, when the system is presented with queries.

The software used to run the simulations described in this paper is Tlearn v1.01 (Plunkett & Elman, 1997). Other simulators were also used and similar results were obtained. The settings used included a learning rate of 0.1, momentum set to 0 (the default) and an initial weight range of -0.5 to 0.5.

Training set 1

In this simulation, the network was trained on five relations that have no overlap. Each relational instance consisted of a unique OBJECT, SUBJECT and RELATION. Within each relational instance, each field or argument was represented by a 1-out-of-5 localist encoding. The first five relational instances (and their associated patterns) that were given to the network to learn are shown in Table 1. Figure 3 (see training set 1) contains a graphic representation of the relational instances and their relationships (or in this particular case, their lack of relatedness).

The system was trained for 20 000 epochs. At around 4000 epochs the error curve smoothed close to zero. In other words, the difference between the target and the actual output values was negligible.

Table 1 — The instances and patterns in training set 1.

Relational instance	*Action*	*Subject*	*Object*
likes(jane, pizza)	10000	10000	00100
buys(fred, book)	01000	01000	00010
hates(mary, dog)	00100	00100	00001
has(gina, icecrm)	00010	00010	10000
sells(bob, hat)	00001	00001	01000

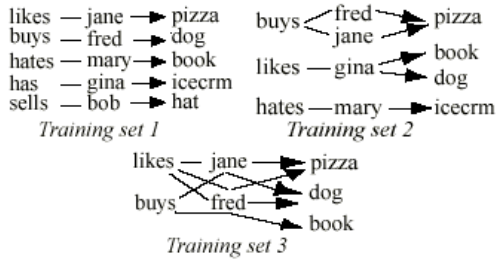


Figure 3 - Graphical representation of 3 training sets.

The system was then presented with a set of test patterns to assess the degree of accessibility that could be obtained (refer to Appendix 1, test pattern set 1). For example, to present the query *likes(jane,X)* the RELATION and SUBJECT input units were set to the patterns for *likes* and *jane* respectively, and the OBJECT input units were set to all zeroes, i.e. the *'X'* is represented by '00000'. Then the OBJECT outputs were inspected. All of the outputs were checked to see if they matched the target values. If the correct number of output units that should be on is N, then a value of 3/(5N) or greater for an output unit is considered to be on, a value of 2/(5N) or less is considered to be off and any values in the region between 2/(5N) and 3/(5N) can be seen to be "partially on". The output thus falls into one of three categories: 1) Either the output is *correct* and all of the outputs units are correctly on or off (as defined above), or 2) The output is *incorrect* and at least one output unit that should be on is clearly off and vice-versa, or lastly, 3) The output is *uncertain* or partially correct, and output units that should be on are only "partially on". For example, if *likes(fred,pizza*) and *likes(fred,dog)* are facts, then if presented with the query *likes(fred,X)*, the answer would be pizza and dog, i.e. N=2. Therefore, the units representing both dog and pizza need to be on and a value of 0.3 (3/(5N)) or greater for both units is needed for the answer to be accepted as correct. Moreover, output units that need to

be off should be less than 2/(5N) or in this case less than 0.2.

With training set 1, correct scores were obtained for all of the test queries. The system was able to handle all of the single and double query test patterns. Therefore, overall an excellent degree of accessibility was achieved with the first training set.

When there is little or no overlap between the elements of the relational instances, the system is able to learn and access elements of the relations with ease. This would correspond with human learning, where the less related information is to other information, the easier it is to understand and learn (Marcus, Cooper & Sweller, 1996; Sweller, 1994).

Training set 2

The next training set the system was given to learn had a higher degree of interaction between the relational instances and their elements than training set 1. In particular, both Fred and Jane like pizza and Jane buys both dogs and books. For the first two instances the same OBJECT is liked by two SUBJECTS and can be characterized as a many-to-one mapping, and for the last two instances the same SUBJECT buys two different OBJECTS, a one-to-many mapping. In contrast, *hates(mary,icecream)* is the only instance that does not overlap with the other four, and is a one-to-one mapping. The OBJECT and SUBJECT in this instance are unique and are not contained in any of the other instances. The five relational instances contained in training set 2 are shown in Table 2. The overlap or interrelations between the elements of the relational instances can be seen graphically in Figure 3 (see training set 2).

The system was trained for 20 000 epochs. At around 2000 epochs the error stabilized at around 0.6 in terms of Tlearn's error measure. The trained net transformed the training patterns as follows:

input		*output*
likes(jane, pizza)	→	likes(fred+jane, pizza)
likes(fred, pizza)	→	likes(fred+jane, pizza)
buys(gina, book)	→	buys(gina, book+dog)
buys(gina, dog)	→	buys(gina, book+dog)
hates(mary, icecrm)	→	hates(mary, icecrm).

It can be seen that the four (related) assertions have now been combined into two assertions. What has been learned is intelligible - *likes(gina,book+dog)* is easy to interpret as signifying that gina likes both books and dogs.

A set of test patterns (see Appendix 1, test pattern set 2) was then given to the system to assess performance on the accessibility property. All output units were then inspected to see if they matched the target units. Using the scoring criterion described above, output patterns were either considered to be 1) correct, 2) incorrect or 3) uncertain. All of the queries with a single unknown element were correctly answered by the system and some of the queries with two unknown elements were correct. None of the queries were considered incorrect, however, four queries obtained uncertain or partially correct scores. These queries were *likes(X,Y), buys(X,Y), X(gina,Y),* and *X(Y,pizza)*. It is interesting to note that all of the responses of questionable correctness need to access information that is contained in more than one relational instance, i.e. information from the many-to-one and one-to-many mappings. For instance, the answer to *likes(X,Y)* is that both Fred and Jane like pizza. This can be clearly expressed in the representation available, but the trained system does not do so. Thus although, accessibility is still relatively good, the net struggles with the queries that access information that has to be integrated from two relational instances.

Table 2 — The instances and patterns in training set 2.

Relational instance	*Action*	*Subject*	*Object*
likes (jane, pizza)	10000	10000	00100
likes (fred, pizza)	10000	01000	00100
hates(mary,icecrm)	00100	00100	10000
buys (gina, book)	01000	00010	00010
buys (gina, dog)	01000	00010	00001

We also tried training the network with the 3 patterns: *likes(fred+jane,pizza*), *buys(gina,book+dog*), and *hates(mary,icecream*), that is, the 3 outputs the net just discussed (call it net 2A) produced in response to the training patterns. The network rapidly learned these patterns, not surprisingly. We tested this network (net 2B) on the queries shown in Appendix 1, test pattern set 2, and found that it had inferior accessibility performance compared with net 2A.

The greater number of uncertain or partially correct scores obtained during testing, for training set 2 (net 2A) reflects the fact that these five assertions may be considered harder to learn. These findings suggest that as the degree of overlap between the relational instances and their elements increases and as the amount of related information that needs to be considered at once increases, so the level of accessibility that the system can cope with, decreases. This corresponds with our understanding of difficulty associated with learning for people. The more interactivity there is between different learning elements, the harder information is to learn (Sweller & Chandler, 1994). The more difficult it is to learn information, the harder it is to transform and use that information. It thus appears, that as the information becomes more complex and so more difficult to learn, the backpropagation system struggles to achieve a

reasonable level of accessibility. The next training set supports this hypothesis.

Training set 3

The last training set has the highest degree of interaction between the relational instances and their elements. The relational instance *likes(fred,pizza)* overlaps with two other instances. The SUBJECT *fred* performs the RELATION *likes* on both the OBJECTS *dog* and *pizza*, a one-to-many mapping. Also, both SUBJECTS *fred* and *jane* perform the RELATION *likes* on the same OBJECT *pizza*, a many-to-one mapping. The five relational instances contained in training set 3 are shown in Table 3. Figure 3 (see training set 3) contains a graphic representation of the relational instances and their interrelatedness.

Table 3 — The instances and patterns in training set 3.

Relational instance	*Action*	*Subject*	*Object*
likes (jane, pizza)	10000	10000	00100
likes (fred, pizza)	10000	01000	00100
likes (fred, dog)	10000	01000	00001
buys (fred, book)	01000	01000	00010
buys (jane, dog)	01000	10000	00001

The system was trained for 20 000 epochs. At around 3000 epochs the error stabilized at around 0.7 in terms of Tlearn's error measure. It should be noted that *buys(fred,book)* and *buys*(*jane,dog)* each have at most one attribute in common with the other instances. The trained net transformed the three more overlapping instances as follows:

input *output*

likes(jane, pizza) → likes(fred+jane, pizza)

likes(fred, pizza) → likes(fred+jane, pizza+dog)

likes(fred, dog) → likes(fred, pizza+dog).

Notice that from this output, there is no way to interpret these instances without inferring that Jane also likes dogs. The whole is not truly equivalent to the sum of the parts, which in this case are the three (and not four) given relational instances. Thus the pattern *likes(fred+jane,pizza+dog)* even if it were valid, would be unintelligible (in contrast to *likes(gina,book+dog)* in training set 2).

The test patterns shown in Appendix 1 (test pattern set 3) were used to test the trained net for accessibility properties. As before, output units were inspected to see if they matched the target units. Using the scoring criterion described above output patterns were either considered to be correct, incorrect or uncertain. All of the queries with a single unknown element were correctly answered by the system. However, only two of the queries with two unknown elements were correct. The two correct queries were *X(Y,dog)* and *X(Y,book)*. The rest of the queries with two unknown elements were incorrect. These queries all access information from more than one relational instance. For example, the query *X(jane,Y)* should have a response of *likes+buys, pizza+dog*. However, the system's response to this query is only *buys, pizza+dog*. As with all the other incorrect queries, some of the relevant information has been lost. It appears that as the information becomes more and more overlapping, the network finds it harder and harder to handle queries that access related elements of information. This type of network appears to be more suited to dealing with one-to-one relations, rather than many-to-many mappings.

Training set 4

A fourth training set was given to the system to learn. It consisted of the relational instances *likes(jane,pizza), likes(fred,pizza), likes(fred,dog), buys(fred,dog),* and *buys(jane,book)*. The amount of overlap between these instances, and the test results, fall somewhere between training sets 2 and 3. All the single unknown element queries were answered correctly. Three of the two unknown element queries were answered correctly, two were uncertain and two were incorrect.

Conclusion

As the degree of overlap between the arguments and predicates of the relational instances in the training set increases, the degree of accessibility provided by the nets simulated decreases. It is well-known that when trained on data that corresponds to a one-to-many mapping, the activations of the output units corresponding to the "many" will be reduced in comparison to a one-to-one mapping. To us, the interesting thing is the effect of argument and predicate overlap on accessibility, and the fact that beyond some critical level of overlap, the trained net starts to produce "generalizations" which, seen from the relational-instance point of view, mean that the net has learnt false propositions e.g. *likes(jane,dog*).

By way of contrast, tensor product networks (Halford et al., 1998) provide full accessibility for arbitrary sets of relational instances, and do not lose critical information when tested.

Backpropagation nets can handle propositional information that is in the form of distinct functions. For example, the model of Rumelhart and Todd (1993) handles propositions such as "canary can fly" in the sense that, given an input "canary can" it produces the output "fly". However, it was not tested for the accessibility property. Our backpropagation net was tested for accessibility, but succeeded in only a limited sense. It could only handle queries to data sets that are relatively simple, in terms of the overlap and relatedness of information. As the relational instances

in the data set become more and more related, so accessibility deteriorates. Consequently the net could not model propositional knowledge adequately. In contrast, a tensor product net can process more complex data sets and still have full access to all the elements of the relational instances.

In a sense, it is not surprising that a backprop-trained net does not do as well at this task - backprop tends to do well at perceptual tasks where generalization of an interpolative type is useful, whereas the data used in this is discrete. Since their introduction, backprop nets and variants have been used in cognitive modeling tasks including those concerned with discrete relational knowledge (Hinton, 1986; Rumelhart & Todd, 1993). This paper has attempted to explore the boundaries of applicability of such models.

What has come out in the wash is evidence from the model's performance of a new dimension of task difficulty. This dimension measures component overlap in a set of facts to be learned. This type of difficulty seems to correlate with model performance at the boundary between rank 1 and rank 2 tasks (in the sense of Halford et al., 1998).

It is clear that humans do have accessibility with respect to their relational knowledge. What might be interesting to investigate is whether they have greater difficulty learning sets of facts like those in training set 3 than those in training set 1, and whether accessibility also takes longer to develop (see Sweller & Chandler, 1994 for a discussion of element interactivity and its effects on learning).

Acknowledgments

This work was supported by a grant from the Australian Research Council. We wish to acknowledge helpful discussions with Steve Phillips, and helpful comments made by a reviewer of a previous version of the paper.

References

Halford, G. S., Wilson, W. H., Guo, J., Gayler, R. W., Wiles, J., Steward, J. E. M. (1994). Connectionist implications for processing capacity limitations in analogies. In K. J. Holyoak & J. Barnden (Eds.), *Advances in connectionist and neural computation theory, vol. 2: Analogical connections.* Norwood, NJ: Ablex.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Brain and Behavioural Sciences,* 21, 803-864.

Hinton, G. E. (1986). Learning distributed representations of concepts. In *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society, 1-12,* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hummel, J. E., & Holyoak, K .J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review,* 104, 427-466.

Marcus, N., Cooper, M., & Sweller, J. (1996). Understanding Instructions. *Journal of Educational Psychology,* 88(1), 49-63.

Plunkett, K., & Elman, J. L. (1997). *Exercises in Rethinking Innateness: A Handbook for Connectionist Simulations.* Cambridge, Mass: MIT Press.

Plate, T. A. (2000). Analogy retrieval and processing with distributed vector representations. *Expert Systems: The International Journal of Knowledge Engineering & Neural Networks,* 17(1), 29-40.

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In D.E. Meyer & S. Korhnblum (Eds*), Attention and Performance XIV* (figure 1.9 p15, top paragraph p16). Cambridge, Mass: MIT Press.

Shastri, L. & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables, and dynamic bindings using temporal synchrony. *Behavioural and Brain Sciences,* 16(3), 417-494.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence,* 46, 159-216.

Sweller, J. (1994). Cognitive load theory, learning difficulty and instructional design. *Learning and Instruction,* 4, 295-312.

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction,* 12, 185-233.

Appendix 1

Test pattern set 1

likes(jane,X), buys(fred,X), hates(mary,X), has(gina,X), sells(bob,X), likes(X,pizza), buys(X,fred), hates(X,dog), has(X,icecrm), sells(X,hat), X(jane,pizza), X(fred,book), X(mary, dog), X(gina, icecrm), X(bob,hat), likes(X,Y), buys(X,Y), hates(X,Y), has(X,Y), sells(X,Y), X(jane,Y), X(fred,Y), X(mary,Y), X(gina,Y), X(bob,Y), X(Y,pizza), X(Y,book), X(Y,dog), X(Y,icecrm), X(Y,hat).

Test pattern set 2

likes(jane,X), likes(fred,X), buys(gina,X), hates(mary,X), likes(X,pizza), buys(X,book), buys(X,dog), hates(X,icecrm), X(jane,pizza), X(fred,pizza), X(gina,book), X(gina,dog), X(mary,icecrm), likes(X,Y), buys(X,Y), hates(X,Y), X(jane,Y), X(fred,Y), X(gina,Y), X(mary,Y), X(Y, pizza), X(Y,book), X(Y,dog), X(Y,icecrm).

Test pattern set 3

likes(jane,X), likes(fred,X), buys(fred,X), buys(jane,X), likes(X,pizza), likes(X,dog), buys(X,book), buys(X,dog), X(jane,pizza), X(fred,pizza), X(fred,dog), X(fred,book), X(jane, dog), likes(X,Y), buys(X,Y), X(jane,Y), X(fred,Y), X(Y,pizza), X(Y,dog), X(Y,book).

What does *he* mean?

Maria Wolters
Rhetorical Systems Ltd.,
4 Buccleuch Place, Edinburgh EH8 9LX, Scotland
maria@rhetoricalsystems.com

David Beaver
Department of Linguistics, Margaret Jacks Hall
Stanford University 94305 Stanford, CA, USA
dib@stanford.edu

Abstract

Empirical results on the meaning of accented pronouns often conflict. This is a problem for formal semantic models. In this paper, we intend to broaden the empirical basis of two of these models. First, in a corpus study, we checked whether properties of the antecedent influence whether a pronoun is be accented. We found that pronouns with NP antecedents are more likely to be accented than those with pronominal antecedents. In a production experiment, we investigate whether accented pronouns signal topic shifts. Although this effect is present in our data, it is very weak. We conclude that a comprehensive model of pronoun accentuation will need to account for the fact that in most cases, that accent appears to be optional. In fact, most instances of accented pronouns in our data could be explained as signs of a rhetorical contrast.

Introduction

In order to interpret a personal pronoun, a listener needs to determine which discourse entity or entities the pronoun specifies. Since pronouns themselves carry very little semantic information, the listener needs to tap into a variety of information sources in order to find out which discourse entity a given pronoun specify: lists of salient discourse entities, grammatical conventions, discourse structure, assumptions about the discourse model of the speaker, and so on.

Formal work on pronoun interpretation has focussed on written language. In this paper, we explore how the insights gained so far can be translated to speech, where phrasing and accentuation might provide important cues to resolution algorithms. Most semanticists take accent on pronouns to signal somewhat "unusual" resolution strategies. In particular, accented pronouns are assumed to signal topic shifts. But do the theories developed so far describe successfully how accented pronouns actually used in speech? This is the question we ask here.

Our paper is organised as follows: First, we sketch the theoretical basis of our analyses, the work of [Nakatani, 1997] and [Kameyama, 1997]. Then, we explore to what extent the patterns postulated by the theories can indeed be found in corpus data and experimental data. Finally, we discuss the consequences of our results for semantic and cognitive approaches to resolving accented pronouns.

Theoretical Claims

It is not clear what exactly speakers mean when they accent a third person personal pronoun. Some people apparently accent a pronoun if the distance in clauses between pronoun and antecedent is larger than usual [Givón, 1983]. A number of sample discourses that have been discussed extensively in the literature [Kameyama, 1997, Beaver, 2000] show that when substituted for their unaccented counterpart in a given discourse, accented pronouns are frequently resolved to a different discourse entity. In the following example, most listeners would resolve the unaccented pronoun in (2) to Ian, whereas the accented pronoun in (3) would be resolved to James.

(1) Ian$_i$ often meets James$_j$ for dinner.
(2) He$_i$ prefers Italian restaurants.
(3) HE$_j$ prefers Italian restaurants.

We will now examine two theoretical analyses of accented pronouns.

Nakatani: Shifts in Attentional State

Christine Nakatani develops a model of stressed pronouns in terms of attentional state [Nakatani, 1997]. Her framework is Centering Theory [Grosz et al., 1995]. The goal of Centering is to develop a theory of *local* discourse structure, i.e. to describe what makes a discourse segment coherent. From a psycholinguistic point of view, Centering models how the attentional state of speaker and hearer change during a discourse. Each attentional state contains a set of discourse entities which are the current "centers" of attention – hence the name "Centering".

In the Centering model, each sentence is associated with a list of the discourse entities which have been realised in that sentence. These discourse entities constitute the list of *forward-looking centers*. The forward-looking centers are ranked according to their salience. The most salient center on the list is called the *preferred* forward-looking center C_p. The most salient entity of the previous utterance U_{n-1} that is realised in the current utterance U_n is the *backward-looking center* C_b. Each sentence has at most one C_b. Transitions between sentences can be classified according to two criteria:

1. whether the backward-looking center of the current sentence is the same as that of the previous sentence

$(C_b(U_n) = C_b(U_{n-1}))$

2. whether the backward-looking center of the current utterance is also the most salient entity of that utterance, i.e. the preferred forward-looking center $(C_b(U_n) = C_p(U_n))$

When the C_b is both maintained and $C_p(U_n)$, we get a *continue* transition. When the C_b is no longer $C_p(U_n)$, the transition is classified as a *retain*, and when the C_b changes, we have a *shift* transition. [Brennan et al., 1987] introduce two kinds of shifts: in *smooth* shifts, the new C_b is also C_p, in *rough* shifts, the new C_b is not the most salient entity in the utterance.

In order to determine the connection between accented pronouns and attentional state, Nakatani analysed a monologue by a gay male native speaker of American English in terms of local and global discourse structure. She found 25 accented subject pronouns; accented object pronouns were even rarer. Of these, 7 occurred when the transition was a smooth shift, and 9 co-occurred with shifts to the backward-looking center of the preceding discourse segment. 6 of the 9 remaining cases were contrastive, 3 required limited inference. Within the Centering framework, accent on a subject pronoun means that the speaker has shifted to a new C_b, to a new "main center of attention". Whether the C_b comes from the same or the preceding segment appears to be irrelevant. The results of [Brennan, 1995] support this analysis. In her spontaneous monologues, pronouns tended to be accented when the antecedent was not the backward-looking center.

Kameyana: Competing Influences

In contrast to Nakatani, [Kameyama, 1997] presents an integrated model of pronoun resolution that draws on several sources. Her approach has the advantage that it sits well with current constraint-based approaches to linguistics such as Optimality Theory [Prince and Smolensky, 1993].

Centering Theory provides the backbone of Kameyama's work: the discourse entities of each utterance are grouped into a partially ordered list, and when a pronoun occurs, it is resolved to the most salient discourse entity in the preceding utterance that is mentioned again in the current utterance. Kameyama calls this discourse entity the *center*. The salience of the discourse entities in an utterance is determined by several potentially conflicting factors. In Kameyama's algorithm for pronoun resolution, the first step is to filter out all referents in the attentional state which violate the highest ranked syntactic and semantic constraints. An example for such a high-ranking syntactic constraint is agreement: masculine personal pronouns in English usually refer to male persons.[1] Then, parallelism, attentional structure (salience), and commonsense inference conspire to yield a basic preference order on the remaining referents which are compatible with the pronoun. The most highly ranked entity on that list becomes *center*, the preferred antecedent for a pronoun in the next sentence.

When the pronoun to be resolved is accented, this reverses the preference order imposed on the list of potential antecedents: the accented pronoun is then taken to refer to the least prominent DR on that list. Note that the order is only reversed *after* high-ranking syntactic and semantic constraints have been considered, so that the antecedents still in the list are indeed viable alternatives. Kameyama thus predicts that accentuation will have no effect when there is only one potential antecedent. She also predicts that stressing the pronoun will not resolve any ambiguity if all potential antecedents are equally salient.

In her paper, Kameyama focusses on two factors that affect salience: the form of the antecedent, which is captured by the *exp order* hierarchy, and the syntactic function of the antecedent, which is captured by the *gr order* hierarchy. On the *exp order* scale, pronouns are more salient than definite NPs, while on the *gr order* scale, subjects are more salient than direct objects, which are in turn more salient than indirect objects or adjuncts. The most salient discourse entity in U_{n-1} is the most likely antecedent for a pronoun in U_n. For example, Kameyama predicts that if a subject pronoun can be resolved to both the subject and the object of the preceding sentence U_{n-1}, and if both subject and object are definite NPs, then the pronoun in U_n will be resolved to the subject of U_{n-1} if it is not accented, and to the object if it is accented, because the subject is more salient than the object.

Bender, Mayer, and Dogil tested that prediction for German [Bender et al., 1996]. They synthesised various short discourses consisting of two sentences. The first sentences had SVO structure; in the second sentence, the subject was pronominalised. In half of the discourses, the second sentence also contained an object pronoun. The subject pronoun was either unaccented or bore one of a number of potential pitch accents. In a 60-minute experiment, listeners were presented with all possible combinations of discourses. For each discourse, they had to indicate the correct interpretation by pointing to a picture. There were four alternatives per discourse. One picture depicted the situation where the pronominal subject of U_n was resolved to the subject of U_{n-1}, one the situation where the pronoun was resolved to the object of U_{n-1}. The other pictures were distractors. They found that the listeners almost never resolved the pronoun to the object NP, even if it was accented. The experiments can be criticised on numerous counts. In particular, the prominence relations between subject and object in the target sentence may not have been realised appropriately [Mayer, 1997]. Still, the experiment demonstrates that the influence of accentuation on pronoun resolution might be more subtle than introspection and corpus studies suggest.

[1] Cases in which that constraint is violated, e.g. females are mistaken for males, are comparatively rare.

Conclusion: Which Evidence is Needed?

The evidence we have surveyed in the preceding paragraphs is conflicting. On one hand, the contexts in which accented pronouns occur do differ markedly from those in which unaccented ones occur. However, it is still not clear what the main function of a pitch accent on a pronoun is. Nakatani's and Brennan's speakers consistently used them to mark what we term "topic shifts"[2] , and Givón's speaker uses accents to mark that the discourse entity the pronoun co-specifies with was last mentioned two or more clauses ago. On the other hand, the listeners of Bender, Dogil, and Mayer apparently did not care whether a pronoun was accented or not, they just went for the default interpretation.

In order to untangle this confusion, we first of all need to supplement the published data with other analyses from a variety of speakers with different socio-cultural backgrounds. In this paper, we report results from two main lines of attack, corpus analyses and production experiments. First, in an analysis of the speech of American radio news readers, we tested some hypotheses that follow from Kameyama's theory. Second, we conducted a production experiment to check to what extent topic shifts influence whether a pronoun will be accented or not.

The Influence of *gr order* on accentuation

Goal of the Study

The goal of this corpus study is to determine some basic conditions under which subject pronouns can be accented. To this end, we examined three speakers from a large corpus of read speech, the Boston University Radio News corpus [Ostendorf et al., 1995]. Since our data comes from read speech, the results may be affected by the speakers' varying ability to read aloud—just as the monologue data is affected by the idiosyncrasies that the speakers exhibit in their speech.

The part of Kameyama's theory that is worked out in detail in [Kameyama, 1997] makes predictions about the accentuation of subject pronouns in cases where the antecedent occurs in the preceding sentence. The main problem with testing her hypotheses on corpus data is that in our radio news texts, almost all potentially ambiguous pronouns are disambiguated semantically by the sentence they occur in. Therefore, we can only examine to what extent violations of the two constraints *gr order* and *exp order* can predict whether a pronoun is accented.

Data

We chose the Boston Radio News Corpus because it is widely available and widely used in the speech community. It provides ToBI-labelled samples of seven professional American newsreaders who worked for the Boston, Mass., radio station WBUR at the time of the recordings. From the corpus, we analysed the prosodically annotated radio stories from three speakers, `f2b`, `m1b` and `m2b`. All speakers write their own copy. Table 1 provides summary statistics about the texts we used.

Table 1: Overview of Radio News Texts

Speaker	`f2b`	`m1b`	`m2b`
No. of Texts	32	9	4
No. of Sequences Analysed	229	45	19
No. of 3rd Sg. Pron. Subj. in U_n	122	22	8
% of These Accented	38.5%	32.8%	0

We look at sequences of two units, U_{n-1} and U_n. Following [Kameyama, 1998], our units are tensed clauses, with one exception: tensed relative clauses that modify NPs are assigned to the unit they occur in. We excluded all sequences of two units where the antecedent of the subject did not occur within the analysis window. This ensures that our analyses is restricted to just those contexts for which the theory we are testing can make any predictions. All sequences that conformed to our criteria were labelled according to four features, *grammatical function* of the antecedent in U_{n-1} (subject, object, adjunct), *form of the antecedent* in U_{n-1} (zero/pronoun/other), *form of the subject* of U_n (zero/pronoun/other), and *accentuation of the subject of* U_n (yes/no). [3] The annotations were performed by one of the authors, a trained linguist, and cross-checked by the other.

Since the brief reports we are dealing with here frequently present conflicting views and opinions on a certain topic, sentence topics are rarely maintained over several units. Secondly, many sentences are constructed according to the pattern "X said that Y", where, strictly speaking, "X said" and "Y" are separate clauses. The reason for this is that the journalistic code of conduct requires journalists to name the source of their information. `f2b`, who writes her own news copy and supplied most of our data, is particularly conscientious in this respect. Therefore, when the current or previous clause is of the form "X said", we extend our window of analysis to include U_{n-2}. Finally, in comparison to spontaneous speech, these speakers "overaccent". This is part and parcel of the distinctive, neutral speaking style is required of news readers. Hence, our results are restricted to a specific communication situation.

Results

The results are summarised in Table 2. Speakers do not tend to accent pronouns when their antecedent is not the subject (Fisher's exact test: $p<0.6$, df=1). Instead, the form of the antecedent, which is covered by Kameyama's *exp order* constraint, exerts a significant influence: whereas 40% of all subject pronouns with NP antecedents are accented, only 18.5% of pronouns with pronominal antecedents carry a pitch accent (Fisher's exact test: $p<0.05$,df=1). This result is surprising, given

[2] Following [Beaver, 2000], we make a terminological shift here and interpret C_b as the *topic* of a sentence.

[3] We did not calculate κ for our annotations since the features are extremely straightforward.

Table 2: Accentuation of 3rd Pers. Sg. Subject Pronouns

	subject	*other*	**total**
pronoun	58%	0.00%	18.5%
not pronoun	41%	22.73%	40%
total	37%	29%	36%

that most formal models operate heavily with salience orderings based on grammatical roles.

A further analysis reveals that although most of these accented pronouns cannot really be analysed as topic shifts, most are involved in some sort of contrast. Take for example text `s14`, read by `f2b`, which is about two democratic contenders for the post of state attorney. It has the highest number of stressed pronouns of all texts, five. In four of these cases, the accented pronoun implies a contrast between the two candidates. Only two of the accented pronouns occur in explicit contrasts. In the two other cases, the accent on the pronoun highlights a potential contrast between the two candidates, the exact nature of which has to be inferred from the text.

Does Accent Signal a Topic Shift?

Kameyama predicts that a subject pronoun in an utterance U_n should be accented if it co-specifies with the object of U_{n-1} only if a higher-ranking discourse entity in U_{n-1} would also be a possible antecedent. If there is only one possible antecedent, there should be no accent on the subject pronoun, because it does not make sense to re-order a list with only one element. On the other hand, Nakatani's data shows that speakers may accent pronouns to mark a topic shift. So, if the main motivation behind accent on pronouns is to signal topic shifts, we would expect speakers to accent pronouns even if they can be resolved unambiguously to the correct antecedent. If accent is a cue to topic shifts in general, we would expect a similar effect when the new topic is expressed by a definite NP.

We examined these questions in a small production experiment. The main hypothesis was that accent is a cue to topic shifts in general: if the subject of U_n co-specifies with the object of U_{n-1}, it is accented, but not when it co-specifies with the subject—even if subject and object differ in gender. We also wanted to find out whether that effect is stronger for pronouns than for full NPs, since the presence of a full NP in subject position is in itself a sufficient cue to a topic shift.

Design

In order to test our hypotheses, we created a set of four-sentence discourses. The first sentence in each discourse, S1, introduces a person P1 in subject position with a proper name. In the second sentence, S2, P1 appears again in subject position, and a second person, P2, is introduced in object position with a proper name. Both persons differ in gender.

The third sentence, S3, is the key sentence. Its structure is varied according to the two variables SHIFT and PRO:

SHIFT: The subject of S3 is either P1, i.e. the subject of S2 (+ SHIFT), or P2, i.e. the object of S2 (- SHIFT)

PRO: The subject of S3 is either a pronoun (+ PRO) or a semantically empty definite NP (- PRO), "the girl" for female, "the guy" for male referents.

Combining these variables yields four experimental conditions: +SHIFT+PRO, +SHIFT-PRO, -SHIFT+PRO, and -SHIFT-PRO

Apart from the subject, no other discourse entities from S2 are mentioned in S3. This way, we avoid potential violations of Centering's definition of C_b, which states that the highest-ranked center of U_{n-1} that is realised in U_n is the C_b of U_n. The final sentence, S4, maintains the subject of S3. The subjects of S2 and S4 are always pronouns. The key sentences S2 and S3 contained no left-dislocated arguments, and the subject in sentences S2 and S3 is always an agent.

We kept the structure of the sentences as simple as possible so that the subjects had less trouble reading them out loud. We also changed the content of S3 and S4 depending on the subject, so that the discourses were both semantically and syntactically unambiguous. There are four possible combinations of conditions, and for each of these combinations, we created three discourses. A sample set of discourses is given in Figure 1.

Method

We divided these discourses into four lists of six discourses each, so that each list contained each condition at least once. Each of these lists was mixed with a list of discourses for two other experiments on intonational meaning and presented to five readers, yielding a total of 20 subjects. We limited the number of discourses per speaker because the experiment was interleaved with two other production experiments and we aimed to keep the total duration of the experiment below twenty minutes, in order not to strain our subjects' voices too much. Because of the small scale of our study, we unfortunately cannot present data on subject-specific variation. All in all, we collected 120 discourses, 30 per condition.

All of our subjects were undergraduates at Stanford University who did not major in Linguistics. One or two of the subjects had a slight cold. Although their mother tongue was American English, they came from different parts of country. They were paid for their participation at a standard rate of $10 per hour. The discourses were recorded in a sound-deadended room. Before the recording was started, the subjects were asked to read through each dialogue. The list of discourses to be analysed later was preceded by four practice discourses. Each discourse was printed in large type on a separate page. The subjects were instructed to read each discourse silently first, make sure they understood what was meant, and then read it out aloud as if they were talking to a friend.

Figure 1: Sample set of four discourses with P1=Julia, P2=Nathan

S1: Julia$_{P1}$ went to a bar last night.
S2: She$_{P1}$ chatted with Nathan$_{P2}$ for a while.

		-SHIFT	+SHIFT
+PRO	S3	**She**$_{P1}$ also spent some time at the bar.	**He**$_{P2}$ had been waiting for this chance for ages.
	S4	Afterwards, she$_{P1}$ was really tired.	He$_{P2}$ had always thought that Julia$_{P1}$ was a nice girl.
-PRO	S3	**The girl**$_{P1}$ also spent some time at the bar.	**The guy**$_{P2}$ had been waiting for this chance for ages.
	S4	Afterwards, she$_{P1}$ was really tired.	He$_{P2}$ had always thought that Julia$_{P1}$ was a nice girl.

None of the subjects reported any difficulties in understanding any of the discourses.

Due to misreadings which were not caught at the time of the recordings, five discourses had to be discarded, which leaves us with a total of 115 discourses. We then randomly discarded another seven discourses, in order to achieve an equal number of samples in each condition. This brings the total number of instances per condition down to 27.

We analysed the discourses both acoustically and perceptually. On the acoustic level, we computed mean, maximum, and range (= maximum F0 - minimum F0) of the logarithm of F0 for all subjects of the third sentence of each discourse. This transformation yields an approximately normal distribution of F0 values. F0 was computed using the `get_f0` program of Entropic ESPS Waves [4] We transformed mean, maximum and range to z-scores based on the mean and standard deviation of a speaker's F0 during the current discourse. These z-scores were then submitted to a statistical analysis. We assume that accentuation will lead to a higher F0 maximum and a higher mean F0. We did not compare durations, since we did not find a satisfactory normalisation procedure.

On the perceptual level, we determined for all discourses whether the subject of S3 was accented, and whether it had been reduced. The definite NPs count as reduced if a speaker has shortened the nucleus of "guy" or "girl"; the pronouns count as reduced if a speaker has replaced the high front vowel of the citation form with a central vowel. These criteria allow for a wide range of dialectal and idiolectal variation in the realisation of the vowels in the target nouns and pronouns. All judgements were made by a trained phonetician; the transcriptions were checked using sonagrams and pitch contours. Table 3 summarises the frequency of accentuations and reductions of each word. The nouns are almost always accented, and rarely reduced. The pronoun "she" is reduced often, sometimes to just the alveopalatal fricative /ʃ/, and rarely accented.

Results

First, we examine the acoustic results. If our main hypothesis is correct, then both nouns and pronouns should be more prosodically prominent when SHIFT is violated.

[4]The target words were almost never spoken with creaky voice, although the register occurs frequently in our data: almost all female speakers use it to mark the end of a sentence.

Table 3: Frequency of accentuation and reduction

	girl	*guy*	**nouns**	*he*	*she*	**pron.**
total	25	29	54	23	31	54
% red.	24%	17%	20%	30%	64%	50%
% acc.	100%	97%	98%	26%	13%	18%

That is, mean and maximum F0 should be higher when the constraint is violated. The raw means presented in Table 4 support this hypothesis for pronouns, but not for nouns. Although the average F0 on pronouns in the +SHIFT-condition is about as high as the average F0 of nouns, this is not necessarily due to a pitch accent. Rather, what we have in many of these cases is a high onset, which is often used to mark paragraph boundaries. For definite NPs, the effect appears to be *reversed*; here, F0 is higher when the topic is maintained than when it shifts.

However, the clear tendencies in Table 4 are not statistically significant. Since the data are not normally distributed, we used the Kruskal-Wallis test to determine whether the presence of a shift significantly affects F0; the number of degrees of freedom is always 1. The test was computed over z-scores, not over absolute values, because z-scores help factor out a large part of the inter-speaker differences. For nouns, there is no effect of a shift, both for the z-scores of mean F0 (Kruskal-Wallis χ^2=0.0037,p<1) and for the z-scores of maximum F0 (K.-W. χ^2=2.01,p-value<0.2). We also do not see an effect for pronouns at the gradient level, neither for mean F0 (K.-W. χ^2=1.77,p<0.2) nor for maximum F0 (K.-W. χ^2=1.70,p<0.2). There is a slight effect on the categorical level: pronouns are more likely to be accented in cases of topic shift (Fisher's exact test, p<0.1, df=1, power 0.84). In the +SHIFT condition, 30% of all pronouns were accented, in the -SHIFT condition, 8%. Although the percentages appear huge, note that the absolute numbers are small, which leads to the realistic level of p<0.1.

Discussion

The patterns we found in the data suggest the hypothesis that speakers may signal topic shifts by accenting pronouns—or at least by making them more prosodically prominent. However, this cue is restricted to pronouns; if a full NP occurs in the subject position, it need not be

Table 4: Effect of conditions on mean and maximum F0 (in Hz; mean standard deviation)

	-SHIFT				+SHIFT			
	mean F0		max F0		mean F0		max F0	
-PRO	180	54	221	65	179	53	211	62
+PRO	173	77	183	78	199	69	212	85

accented, because the switch from pronoun to full NP is enough.

There are several reasons why these patterns did not turn out to be significant. For one, we need to control the potential emphatic or contrastive foci in S3 more stringently. For example, in the sentence "X had to explain all the algorithms twice." almost all of our subjects stressed "all", "algorithms", and "twice". But the main problem was that speaking styles vary considerably. Some speakers produced very stereotypical, monotonous intonation contours, while others read the discourses almost naturally. Therefore, we will switch to a design based on spontaneous speech in follow-up experiments, which will also incorporate a dedicated speaker factor. The power of the statistical tests, which was between 0.3 and 0.4, will also need to be increased.

Conclusion

What does it mean to accent a pronoun? The data we have examined, both in the corpus study and the production experiment, do not support either Nakatani's or Kameyama's models of pronominal accent. Thus our studies underline the need for further theoretical and empirical work in this area.

In our data, two trends can be discerned: Firstly, any formal semantic theory of accented pronouns needs to deal with the fact that in many cases, this accent may be *optional*. To what extent speaking style influences whether speakers will choose to accent a pronoun, and to what extent accented pronouns aid comprehension, needs to be investigated by future experiments. Secondly, most of the accented pronouns in our corpus data could be interpreted as cues to some sort of contrast. Often, that contrast was not directly obvious from the context; Listeners need to construct a contrast based on their interpretation of a text to accomodate the presence of the accent on the pronoun. We are currently working on an optimality theoretic account of stressed pronouns that incorporates this finding.

Acknowledgments

We would like to thank Amanda Wildner for help with the measurements and Susan Brennan and an anonymous reviewer for their comments. M.W. conducted most of the work while still at the University of Bonn. D.B. would like to thank the Stanford University Office of Technology Licensing for funding.

References

[Beaver, 2000] Beaver, D. (2000). Centering in optimal theory. Department of Linguistics, Stanford University.

[Bender et al., 1996] Bender, A., Dogil, G., and Mayer, J. (1996). Prosodic disambiguation of anaphoric pronouns in german discourses. In *UCREL Technical Papers, Vol.8 – Special Issue: Approaches to Discourse Anaphora: Proceedings of DAARC-96, Lancaster, July 17-18, 1996*, pages 28–39.

[Brennan, 1995] Brennan, S. E. (1995). Centering attention in discourse. *Language and Cognitive Processes*, 10(2):137–167.

[Brennan et al., 1987] Brennan, S. E., Friedman, M. W., and Pollard, C. J. (1987). A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics,* Stanford, CA, 6–9 July 1987, pages 155–162.

[Givón, 1983] Givón, T. (1983). Topic continuity in spoken english. In Givón, T., editor, *Topic Continuity in Discourse*, pages 343–364, Amsterdam; Philadelphia, PA. John Benjamins.

[Grosz et al., 1995] Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

[Kameyama, 1997] Kameyama, M. (1997). Stressed and unstressed pronouns: Complementary preferences. In Bosch, P. and van der Sandt, R., editors, *Focus: Linguistic, Cognitive and Computational Perspectives*. Cambridge University Press.

[Kameyama, 1998] Kameyama, M. (1998). Intrasentential centering: A case study. In Walker, M. A., Joshi, A. K., and Prince, E., editors, *Centering Theory in Discourse*, pages 89–112. Oxford University Press.

[Mayer, 1997] Mayer, J. (1997). Intonation und Bedeutung. Arbeitspapiere des Instituts für maschinelle Sprachverabeitung, Stuttgart 4(3), Institut fr Maschinelle Sprachverarbeitung, Universität Stuttgart.

[Nakatani, 1997] Nakatani, C. (1997). *The Computational Processing of Intonational Prominence: A Functional Prosody Perspective*. PhD thesis, Harvard University.

[Ostendorf et al., 1995] Ostendorf, M., Price, P., and Shattuck-Hufnagel, S. (1995). The Boston University radio news corpus. Technical report, Boston University.

[Prince and Smolensky, 1993] Prince, A. and Smolensky, P. (1993). Optimality theory: Constraint interaction in generative grammar. Technical Report 2, Rutgers University Center for Cognitive Science RuCCS.

Structural Determinants of Counterfactual Reasoning

Daniel Yarlett (dany@cogsci.ed.ac.uk)
School of Cognitive Science, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW Scotland

Michael Ramscar (michael@dai.ed.ac.uk)
School of Cognitive Science, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW Scotland

Abstract

In this paper we explore the hypothesis that the processes underlying the matching and use of knowledge during counterfactual reasoning are the same as those that underlie reasoning by analogy. We report two experiments which indicate that, in common with analogical reasoning, counterfactual reasoning exhibits the following properties: (i) it is sensitive to systematic structural congruencies between representations; (ii) when background knowledge has to be retrieved before it can be exploited, the impact of structurally congruent background knowledge on the counterfactual inferencing process is mediated by featural commonalities.

Introduction

If it had not been for a young student at the University of Chicago in 1960, the Cuban missile crisis might have escalated to a nuclear war. At first blush this counterfactual sounds absurd, a *non sequitur* – how can there possibly be a link between a single student and the avoidance of nuclear war? It turns out, though, that the assertion isn't quite as far fetched as one might initially think:

> "Not yet twenty-one and too young to vote, the student worked in the Kennedy campaign. He was asked by the local Democratic organization if he would vote on behalf of a dead voter whose name was still on the rolls. He readily agreed and, refusing the small remuneration that was offered, forged the dead voter's signature and voted a straight Democratic ticket. The Illinois vote was close – Kennedy took the state by fewer than 10,000 votes – and critical. Illinois gave Kennedy the necessary electoral votes to win the presidency." (Lebow and Stein, 1996, p.119).

This additional knowledge connecting the student's actions to Kennedy's electoral success allows the rest of the story to fall into place. Had the student – and, crucially, others like him – not forged votes, Nixon may have been elected instead of Kennedy; if Nixon had been elected he would have established a much less liberal administration; a less liberal administration would have responded more forcefully to Khruschev's deployment of missiles in Cuba; and the odds are that this would have led to further military escalation and, ultimately, nuclear war.

This is not a watertight argument of course – the strength of some of the links in the argumentative chain is questionable – but it does serve to make apparent the relation between the two events which at first appeared utterly disparate.[1] And understanding that there is a connection between the existence of a group of students and the avoidance of nuclear war makes the initial counterfactual seem much more plausible than it did before: if those students hadn't existed then we can now see that this might just have tipped the balance in the favour of nuclear war. Background knowledge is clearly crucial to our assessment of counterfactuals, then, but how it is matched and used during counterfactual reasoning?

In this paper we explore the possibility that the processes underlying the matching and use of knowledge during counterfactual reasoning are the same as those that underlie reasoning by analogy. Although there has been extensive research into counterfactual reasoning over the last two decades from a cognitive-functional perspective (for representative overviews see Roese, 1997; Kahneman & Miller, 1986; Byrne & Tasso, 1999), no-one has explicitly investigated the effect of structural congruency on the inferencing process. We report two novel findings in support of the hypothesis that analogical mechanisms do underpin counterfactual reasoning in at least some contexts.

In Experiment 1 we show that an analogical match between background knowledge and a scenario about which subjects have to reason counterfactually significantly boosts their assessments of the soundness of related counterfactual inferences. This suggests that when engaged in counterfactual reasoning subjects rely on systematic structural matches between representations just as they do when making analogical inferences (Gentner, 1983; Holyoak and Thagard, 1995). Furthermore, previous studies in analogical inference and problem solving have found that it is difficult to exploit the information contained in analogies during real-life problem-solving because analogies are typically difficult to retrieve (Gick and Holyoak, 1980; Gentner, Ratterman and Forbus, 1993). In Experiment 2 we add a retrieval requirement to the task investigated in Experiment 1, and show that this same finding also applies to counterfactual reasoning.

Why Analogy?

Theories of counterfactual reasoning agree that similarity plays a core role in determining legitimate inferences – in order to evaluate what would be true if A were the case, it makes sense to consider only the state of affairs that is most similar to the way things actually are, except that in

[1] See Lebow and Stein (1996) for a fuller analysis of the plausibility of this counterfactual sequence of events.

it *A* is true instead of false. This similarity constraint prevents unnecessary or extraneous alterations to the situation under consideration that could otherwise bias the inferencing process, and is what makes counterfactual reasoning distinct from the idle entertainment of hypothetical situations. For example, consider the relatively uncontentious counterfactual 'If this match had been struck, it would have lit'. As it stands this is most likely a true thing to say (depending on the precise context of utterance), but without the similarity constraint one can trivially refute it by making unwarranted changes to the basic situation and hence considering situations where matches are non-flammable, or where all matches are underwater, and so on.

However, despite the central role that similarity plays in theories of counterfactual reasoning, no consensus exists on how it should be defined. Possible world theorists treat it as a primitive partial ordering between possible worlds (Lewis, 1973; Stalnaker, 1968); law-based theorists treat it as the preservation of consequences of natural laws (Chisholm, 1946; Goodman, 1947; Pollock, 1976); whilst others treat it as minimal adjustments to autonomous causal mechanisms represented in dependency structures (Pearl, 2000). These formal characterisations of similarity and the constructs they rely on make no claim to being cognitively plausible, and are furthermore open to charges of underspecification (Lewis, 1973, §4.2; Pollock, 1976, p.17; Fine, 1975; and Bowie, 1979).

In contrast, work on the notion of similarity in cognitive science has concentrated on the empirical testing of concrete proposals about what it means for two things to be similar to one another. One finding is that the shared and distinct features of two representations play an important role in determining their similarity (Tversky, 1977), but more recent work has also shown that the *relational structure* holding between features also affects judgements of similarity (Gentner & Markman, 1997; Love, 2000).

Gentner and Markman (1997) point out the effect of structure on similarity judgements is consistent with the idea that analogical alignment plays a role in the evaluation of similarity. If analogy really does play a role in determining similarity then, given that similarity is so central to counterfactual reasoning, it makes sense to investigate whether analogy plays a role in counterfactual reasoning by testing for effects of structure on the inferencing process. Furthermore, the account of similarity provided by the analogy-based *structural alignment* framework (Gentner, 1983) has several merits from a cognitive perspective.

Firstly, the structural alignment framework has been subject to a program of empirical testing which provides evidence regarding its plausibility as an account of the way that people compare representations (see Gentner & Markman, 1997, for a review). Secondly, the structural alignment process can be modelled computationally (e.g. SME; see Falkenhainer, Forbus and Gentner, 1989) meaning that the processes it relies upon are tractable. The final reason for preferring the structural alignment account of similarity over alternative accounts is that the analogical mechanisms that it is founded upon have been explicitly connected with many other cognitive skills – most relevantly inference (see Holyoak and Thagard, 1995, for a good review of the many applications of analogy) – offering yet further reason to believe that it may also be implicated in the process of counterfactual reasoning at the cognitive level.

	+ Structure	− Structure
+ Features	Literal Similarity	Surface Similarity
− Features	Analogy	First-Order Relations

Table 1: The commonalities each variant category shares with its base in the materials used by Gentner, Ratterman & Forbus (1993).

Given these reasons, the question of whether the same empirical findings about analogical inference can also be identified in counterfactual reasoning tasks is a pertinent one. If commonalities between counterfactual and analogical reasoning could be established empirically, this would undoubtedly be profitable to our understanding of counterfactuals due to the mature status of research into theories and models of analogy.

Previous Work on Analogical Matching

Experiments conducted by Gentner, Ratterman and Forbus (1993) investigated the effect that systematically varying featural and structural matches between two representations of knowledge had on three measures: the *retrievability* of one item given the other as a prime; on assessments of *inferential soundness*[2] between the two items; and on the perceived *similarity* of the two items. Their series of studies used materials consisting of a *base* scenario and four *variants* of this base in the 2×2 design depicted in Table 1. The systematic variation of featural and structural matches in this design allowed the influence of both factors on a range of tasks to be assessed, and results in a taxonomy of four different types of variants to a base: literal similarity (LS), surface similarity (SS), analogy (AN) and first-order relations (FO).

An important finding reported in this series of studies was that the inferential soundness of two domains is determined primarily by the sort of systematic structural congruencies that analogy is sensitive to. Experiment 1 tests whether this result applies to counterfactual reasoning too, by measuring whether analogical matches to background knowledge boost evaluations of the soundness of counterfactual inferences in the same way as they do in analogical reasoning.

An additional finding about analogical inference is that structural matches can only be exploited in the inferencing process if they are first retrieved, which is typically a difficult thing for people to do (Gick and Holyoak, 1980). This suggests another hypothesis that can be tested: if the cognitive processes underlying counterfactual reasoning are really like those underlying analogy, then in contexts that demand retrieval we would expect analogy based inferences to occur relatively infrequently, because the impact of structural matches on the counterfactual inferencing process is mediated by the featural similarities that drive retrieval (Gentner, Ratterman and Forbus, 1993; Ramscar and Yarlett, 2000). This prediction is tested in Experiment 2.

[2]Inferential soundness is defined as the degree to which knowledge about one domain can be used to generate appropriate inferences about another domain.

Experiment 1

Analogy is universally held to be determined by systematic congruencies in the relational structure of two representations. If analogical mechanisms really do underpin counterfactual reasoning then we would expect to find that structural matches support counterfactual reasoning in exactly the same way that they support analogically based reasoning (Falkenhainer, Forbus & Gentner, 1989; Gentner, Ratterman & Forbus, 1993). Moreover, we would expect a simple featural match between a counterfactual reasoning problem and background knowledge to have no such facilitatory effect.

The simplest test for an effect of shared structure as opposed to shared features in counterfactual inference, as predicted by the hypothesis that analogical mechanisms underpin counterfactual reasoning, is to place subjects in two conditions: one in which they have information that counterfactually matches the problem scenario structurally but not featurally (AN condition); and the other in which the opposite is the case (SS condition). The effect of the knowledge on the soundness ratings assigned a counterfactual inference related to the two sets of background knowledge can then be measured relative to a pre-test condition (the *a priori* or AP condition) in which subjects are provided with *no* background knowledge relevant to the counterfactual in question. If counterfactual reasoning really is sensitive to analogical matches then we would expect a significant increase from the pre-test soundness ratings in the AN condition, but no such increase in the SS condition.

Method

Subjects read either the SS or AN match to a base scenario, and were then asked two comprehension questions about it to ensure they had read it properly. They were then asked to read the corresponding base, and reason counterfactually about it: subjects had to rate the soundness of four counterfactual inferences which undid a key event described in the base. Both SS and AN variants were designed to promote the same counterfactual inference by featuring it as the outcome of the sequence of events they described. The relative effect of a featural versus a structural match between the base problem and background knowledge on the inferencing task could therefore be measured by comparing the soundness ratings for the same inference in each condition. Soundness ratings were elicited by asking subjects how likely the counterfactual inference was to be true. Ratings were provided on a 9 point scale anchored by 'extremely unlikely' at the lower end, and 'extremely likely' at the upper end.

Participants Participants were 40 undergraduate students of the University of Edinburgh, completing an introductory course in psychology. All participants were volunteers, and no reward was offered for taking part in the experiment.

Materials The materials were designed to replicate the experimental design used in Gentner, Ratterman and Forbus (1993) as summarised in Table 1, while also being compatible with a counterfactual reasoning task. 5 base descriptions and 4 variant categories for each base (corresponding to the LS, SS, AN, FO categories described earlier) were designed, resulting in 25 scenarios in total. Only the SS and AN variants were used in Experiment 1.

> Ostavia, Grern and Donnol were three neighbouring and hostile states that seemed to take a particular delight in antagonising one another. It wasn't uncommon these days to hear of one country criticising the others, and threatening them with military action. The trouble had begun several decades ago when Ostavia had attacked both Grern and Donnol in a failed attempt to invade them. Since then relations between the three states had gone downhill, and just recently things appeared to be getting even worse.
>
> **Ostavia had recently launched an attack against Grern**, and although resistance was fierce it looked as though Ostavia was soon going to conquer the Grern forces. However, Grern approached Donnol for help in the conflict, arguing that if they didn't prevent the onslaught then Ostavia would be coming for them next. Donnol considered Grern's argument and saw that its conclusion was probably correct. Therefore Donnol sent its troops to reinforce Grern's units, and after an intense struggle the Ostavian attack was repelled. *Soon after, Grern and Donnol formed a long-lasting alliance, and occupied Ostavia enforcing harsh sanctions on its citizens.*

Figure 1: Example base description. The key event for this description is shown in bold type, and the outcome altered in the variant conditions in italics (distinctions were unmarked in experimental materials).

Each base was structured so that it described a key event which led to a specific outcome (see Figure 1 for an example). For each base, four alternative counterfactual outcomes to the actual outcome described in the base were selected (these were the alternatives rated for soundness by subjects). The counterfactual analogy (AN) variant to the base in Figure 1 is shown in Figure 2. The counterfactual inferences corresponding to the example base scenario that subjects had to rate for soundness are also shown, in Figure 3.

Each of the four variants of a particular base was designed to undo the key event described in the base while either sharing or not sharing featural and structural commonalities with it, depending on which of the variant categories it belonged to. All variants in a material set were constructed to end in an identical outcome, different from that occurring in the base of the set (in the example materials shown the four variants ended with some form of reconciliation between the parties involved, instead of the 'harsh sanctions' mentioned in the base). The four variants can thus be regarded as providing different types of background knowledge, each of which could serve to promote the plausibility of the same counterfactual inference. The issue of interest in this paper is how these four categories of background knowledge, and the different ways that they are related to the base scenario, affect subjects' soundness judgements about the counterfactual outcomes they describe.

The four variants to a base were created in the following manner. First, a key event in each base was identified. In those variants to a base in which a counterpart to the key event should be easily discernible (i.e. in the LS and AN structurally congruent variants), the appropriate counterpart was explicitly negated to make it counterfactual to the base with respect to the key event. In those variants in which a counterpart of the key event should not be easily identifiable

A chess-playing craze had recently swept through the inhabitants of Chesterton, and as a result three chess clubs had sprung up in the sleepy town over the last few years. The clubs were, unimaginatively enough, called the Maters, the Gambit Players and the Rank and Filers. Times were hard for the clubs, because there simply weren't enough players to go around. The problems had been started by the antagonistic members of the Maters. Several months earlier the Maters had scandalously offered reduced subscription rates to members of the Gambit Players and Rank and Filers. Since then an air of hostility had descended upon the chess-playing community in Chesterton.

The Maters thought about trying again to gain extra members from the other two clubs, this time by promising them extra facilities as well as reduced subscription rates. **However, eventually the Maters thought this was a bad idea because it seemed too underhand and risky.** Unbeknownst to the Maters, the other two clubs had formed a pact with one another that they would merge and drive the Maters out of town if they tried to steal any of their members again. *However, because the Maters thought better of their plan everyone relaxed a little, and a couple of years later the clubs were on very friendly terms.*

Figure 2: Example analogy (AN) variant. The counterfactual counterpart of the key event is shown in bold type, and the counterfactual outcome is shown in italics (distinctions were unmarked in experimental materials).

If Ostavia hadn't decided to launch an attack against Grern and Donnol, it wouldn't have been defeated in war *and something else would probably have happened*. Rate the following outcomes according to how likely you think they would have been to occur if Ostavia hadn't launched its attack (please circle *one* number for each of the four options):

1. **All three countries would have ended up disarming, and developing friendly relations.**
2. The countries would have organised a yearly carnival in celebration of newfound connections.
3. The countries would have merged under an even greater threat.
4. The countries would not have learnt from their mistakes, and would end up fighting one another again.

Figure 3: Example counterfactual outcomes that subjects were asked to rate for soundness. The inference promoted by the base's four variants is shown in bold type.

(i.e. in the SS and FO conditions, where there is no structural congruity with the base), no event was described that could be construed as being an instance of the key event.

Procedure Subjects were exposed to four of the material sets, two in the SS condition and two in the AN condition. Materials were randomised across subjects, and order of presentation was counterbalanced.

The written instructions to the experiment informed participants that they were taking part in a study investigating general reasoning, and that they were going to be shown a series of scenarios and asked to make some inferences about them. Participants were further told that for each problem they were going to be provided with some background information (this information corresponding to the SS and AN conditions of the experiment), and then asked to reason about a problem scenario. They were explicitly informed that the background information **may** be of use, and that they should only make use of it if they thought it was relevant to the problem in hand.

Pre-test

A pre-test condition was run before Experiment 1 to allow the facilitatory effects of providing various categories of background knowledge to be compared to the case in which subjects are provided with *no* relevant background knowledge. 20 postgraduate students of the Division of Informatics, University of Edinburgh, were provided with the 5 base descriptions in a semi-randomised order, and asked to evaluate how sound it seemed that each of 4 counterfactual alternatives would occur had the key event in each description been undone. Ratings were elicited on a 9 point scale. Two sets of booklets presented all 5 base descriptions and the corresponding counterfactual inferences to be rated in reverse orders to minimise order effects. Subjects were randomly assigned to either of the two booklet conditions.

Results

The mean soundness ratings of the promoted inferences in the SS- and AN-exposure conditions from Experiment 1, along with the pre-test (AP) soundness ratings for the same inferences, are shown in Figure 4. T-testing revealed no significant shift in soundness ratings from the pre-test to the SS condition (Welch's $t = 0.36$, $df = 58.05$, $p > 0.025$, one-tailed). In contrast, soundness ratings did increase significantly from the pre-test to the AN condition (Welch's $t = 2.00$, $df = 58.64$, $p < 0.025$, one-tailed).[3]

Discussion

The results of this experiment indicate that subjects will reason counterfactually on the basis of an analogical match to background knowledge when one is available, and that a featural match to background knowledge alone is insufficient to support a counterfactual inference. The fact that an analogical match to background knowledge makes a counterfactual inference seem more sound shows that people are sensitive to the sort of systematic structural correspondences found in analogy when engaged in this form of reasoning, and therefore supports the hypothesis that analogical mechanisms underpin counterfactual reasoning at the cognitive level.

Experiment 2

Gentner, Ratterman and Forbus (1993) present evidence that in contexts where subjects first had to retrieve knowledge in order to be able to exploit it to make an analogical inference

[3]Both tests involved samples with unequal variances by Levene's test, and hence Welch's corrected t statistic was calculated. Tests were conducted at adjusted α levels to keep the overall risk of a Type I error at the 0.05 level, by solving $1-(1-\alpha)^2 = 0.05$ for α.

(i.e. where relevant knowledge is not presented in association with the base it pertains to as it was in Experiment 1), both structural *and* featural matches play a role in determining the usefulness of that knowledge in the inferencing process. The effect of featural commonalities is introduced because the process of retrieval is sensitive to featural, but not structural, matches (Gentner, Ratterman and Forbus, 1993; Ramscar and Yarlett, 2000). Therefore, if counterfactual reasoning really is similar to analogy at the cognitive level, we would expect the addition of a retrieval requirement to the inferencing task of Experiment 1 to have the following effects.

First, the AN condition should no longer produce a significant boost in soundness ratings compared to the pre-test condition. We expect this because AN matches are typically difficult to retrieve (see, for example, Gick and Holyoak, 1980), and are therefore less likely to be used in the counterfactual inferencing process in this context. Second, the soundness ratings of SS condition inferences should remain the same as in Experiment 1 (because SS materials have high retrievability the additional retrieval requirement should make no difference to the availability of this knowledge in the inferencing process). Third, and finally, the overall pattern of results over the 4 variant categories should exhibit sensitivity to both featural and structural factors, instead of just structural factors as in Experiment 1 (because featural matches facilitate retrieval, which mediates the influence of structure on the inferencing process).

Experiment 2 was accordingly designed to test these hypotheses by adding a retrieval task to Experiment 1, so that background knowledge had to be successfully retrieved before it could be used as the basis for supporting a counterfactual inference.

Method

In the first phase of the experiment subjects were asked to read 5 variants, one from each material set. Two multiple-choice comprehension questions were asked after each story had been read, in an attempt to ensure that subjects read the information thoroughly. They were next asked to complete one of two short distraction tasks. The distraction tasks were unrelated to the present study; one investigated a spatio-temporal priming phenomenon, and the other morphological inflection. In the second phase of the experiment subjects were asked to perform exactly the same inferencing task as described in Experiment 1, except that this time they had also been provided with the LS and FO variants of background knowledge, in addition to the SS and AN categories. However, because the connection between the base problems and the background knowledge they were provided with in the variants was not made explicit, subjects had to retrieve relevant knowledge before being able to use it in the inferencing process. Subjects were not told that there was any connection between the first and second part of the experiment. This was done to determine whether subjects would be able to *spontaneously* recruit (Kahneman and Miller, 1986) background knowledge in order to make their soundness judgements for the counterfactuals.

Participants Participants in this study were 148 undergraduates and postgraduates at the University of Edinburgh. All participants were volunteers. The undergraduate students were awarded class credit for taking part in the study.

Materials All 5 bases, with their corresponding 4 variants (LS, SS, AN and FO categories), of the materials described in Experiment 1 were used.

Procedure 1 variant was taken from each of the 5 material sets, to result in each prime set. This meant that there were 4 prime sets in total. Two random orderings of each prime set were combined with two random orderings of the question set (which was the same as that used in the pre-test and Experiment 1), to create four distinct material orderings for each of the four prime sets. This was done to minimise the potential for presentation effects, and resulted in 16 types of booklet in total. The 2 distractor tasks were randomly included in the booklets, and took subjects less than five minutes to complete.

In the first part of the experiment subjects were instructed only that they were taking part in a memory experiment, and that they should therefore read the stories they were presented with carefully. In the second part of the experiment no reference was made to the stories in the first part; subjects were merely asked to provide soundness ratings as they saw fit.

Results

The results of Experiment 2 are shown in Figure 4 (along with the results from Experiment 1).

A 2×2 ANOVA was conducted on the four experimental conditions. This revealed a significant effect of featural matches ($F(1,147) = 8.75, p < 0.01$), a significant effect of structural matches ($F(1,147) = 16.76, p < 0.01$), and no significant interaction between the two factors ($F(1,147) = 3.72, p > 0.05$) consistent with the predictions of the analogy-based account of counterfactual reasoning.

The soundness ratings in neither the SS (Welch's $t = 0.81$, $df = 49.03$, $p > 0.025$, one-tailed) nor the AN (Welch's $t = 0.90$, $df = 51.61$, $p > 0.025$, one-tailed) conditions differed significantly from those in the pre-test.[4] Whilst the effect of SS knowledge on the counterfactual inferencing process is unchanged between Experiments 1 and 2, the facilitation of soundness ratings produced by the analogical knowledge observed in Experiment 1 has disappeared. The best explanation of this difference is that the retrieval requirement introduced in Experiment 2 has prevented AN matches from being successfully retrieved in the counterfactual inferencing process, a finding paralleled in analogy research.

Discussion

The results of Experiment 2 were consistent with those predicted by the analogy-based account of counterfactual reasoning. Whereas the provision of analogical knowledge in Experiment 1 was sufficient to boost the soundness of related counterfactual inferences, the same effect of analogical knowledge was not observed in Experiment 2. The best explanation of this seems to be that retrieval does appear to mediate the effect of structural congruency in the solution of

[4] Corrected α values were used as before. Levene's test indicated both t-tests involved samples with unequal variances, and so Welch's corrected t was calculated.

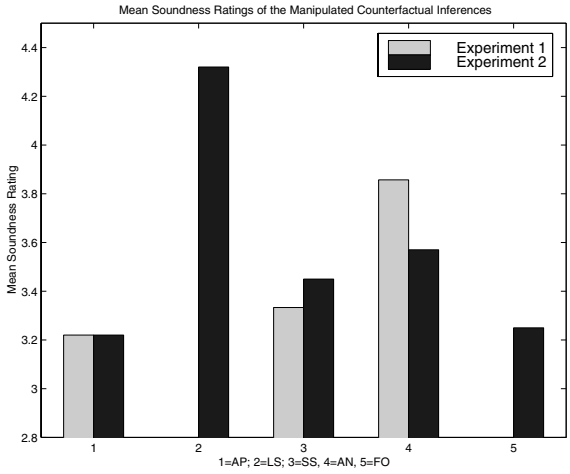


Figure 4: A bar chart plot of the mean plausibility of the manipulated counterfactual inferences in the pre-test and Experiments 1 and 2.

the counterfactual reasoning problems. Furthermore, both featural and structural factors were shown to have a significant effect on the soundness ratings provided by subjects, suggesting that both a feature sensitive retrieval process and structural alignment have a role to play in accounting for the way in which background knowledge is exploited during counterfactual reasoning. This pattern is in accordance with the research on analogical reasoning reported by Gentner, Ratterman and Forbus (1993).

General Discussion

We have presented evidence that there are deep similarities in the ways in which knowledge is exploited during both analogical and counterfactual reasoning. Specifically, we have shown that an analogical match to stored knowledge of an appropriate kind is sufficient to significantly boost the plausibility of a counterfactual inference, and that the structure-matching process that we hypothesise underpins counterfactual inference is mediated by a retrieval process sensitive to featural-matches, just as in analogical reasoning. The former finding is most significant, perhaps, because it is unpredicted by current theories of counterfactual reasoning.

Whilst there are limitations to the current work – chiefly that it doesn't explore what occurs when complete matches to stored counterfactual knowledge are unavailable, and counterfactual representations have to be *constructed* – we nevertheless feel that the current research presents the intriguing possibility of applying existing findings about to counterfactual reasoning, in order to improve our understanding of this important cognitive process.

Acknowledgements

We would like to thank Scott McDonald, Keith Stenning, Lera Boroditsky and the three anonymous reviewers for useful comments on the work reported in this paper.

References

Bowie G.L. (1979). The Similarity Approach to Counterfactuals: Some Problems. *Noûs*, 13, 477-498.

Byrne R.M.J. and Tasso A. (1999). Deductive Reasoning with Factual, Possible, and Counterfactual Conditionals. *Memory & Cognition*, 27(4), 726-740.

Chisholm R.M. (1946). The Contrary-to-Fact Conditional. *Mind*, 55(220), 289-307.

Falkenhainer B., Forbus K.D. and Gentner D. (1989). The Structure-Mapping Engine: Algorithm and Examples. *Artificial Intelligence*, 41, 1-63.

Fine K. (1975). Critical Notice of *Counterfactuals*. *Mind*, 84, 451-458.

Gentner D. (1983). Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, 7, 155-170.

Gentner D., Ratterman M.J. and Forbus K.D. (1993). The Roles of Similarity in Transfer: Separating Retrievability from Inferential Soundness. *Cognitive Psychology*, 25, 524-575.

Gentner D. and Markman A.B. (1997). Structural Alignment in Analogy and Similarity. *American Psychologist*, 52(1), 45-56.

Gick and Holyoak (1980). Analogical Problem Solving. *Cognitive Psychology*, 12, 306-355.

Goodman N. (1947). The Problem of Counterfactual Conditionals. *The Journal of Philosophy*, 44(5), 113-128.

Holyoak K.J. and Thagard P. (1995). Mental Leaps: Analogy in Mental Thought. MIT Press, Cambridge, Massachusetts.

Kahneman D. and Miller D.T. (1986). Norm Theory: Comparing Reality to its Alternatives. *Psychological Review*, 93(2), 136-153.

Lebow R.N. and Stein J.G. (1996). Counterfactuals and the Cuban Missile Crisis. In Tetlock P.E. and Belkin A. (eds.), *Counterfactual Thought Experiments and World Politics: Logical, Methodological and Psychological Perspectives*, Chapter 5, p.119-148, Princeton University Press, Princeton, New Jersey.

Lewis D.K. (1973). Counterfactuals. Harvard University Press, Cambridge, Massachusetts.

Love B. (2000). A Computational Level Theory of Similarity. *Proceedings of the 22nd Annual Meeting of the Cognitive Science Conference*, 316-321.

Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge.

Pollock J.L. (1976). Subjunctive Reasoning. Volume 8 of *Philosophical Studies Series in Philosophy*, D. Reidel Publishing Company, Dordrecht, Holland.

Ramscar M.J.A. and Yarlett D.G. (2000). A High-Dimensional Model of Retrieval in Analogy and Similarity-Based Transfer. *Proceedings of the 22nd Annual Meeting of the Cognitive Science Conference*, 381-386.

Roese N.J. (1997). Counterfactual Thinking. *Psychological Bulletin*, 121, 133-148.

Stalnaker, R. (1968). A Theory of Conditionals. In Rescher, N. (ed.) Studies in logical theory, number 2 in American Philosophical Quarterly Monograph Series, Blackwell Press, Oxford.

Tversky, A. (1977). Features of Similarity. *Psychological Review*, 84(4), 327-352.

Competition between linguistic cues and perceptual cues in children's categorization: English- and Japanese-speaking children

Hanako Yoshida (hayoshid@indiana.edu)
Linda B. Smith (smith4@indiana.edu)
Cindy Drake (cdrake@indiana.edu)
Joy Swanson (jomswans@indiana.edu)
Leanna Gudel (lgudel@indiana.edu)
Department of Psychology Indiana University, 1101 E. 10th Street
Bloomington, IN 47405-7007 USA

Abstract

Past research on children's categorizations has centered on the question of whether categorization is based primarily and /or strictly on perceptual properties or on more conceptual notions of kinds. This paper reports new data pertinent to this issue by examining the influence of linguistic cues and perceptual cues. The results showed that the linguistic cues strongly influenced English-and Japanese-speaking children's judgments in systematic manner. This suggest that for both language groups, linguistic cues activated different conceptual understanding that direct children's attention to different properties. Nevertheless, the linguistic cues had their effect only given named things with ambiguous properties.

Introduction

Children categorize the world systematically. The systematicity of early categories has been well documented by many researchers (Mandler, Bauer and McDonough, 1991; Landau, Smith and Jones, 1998a, 1998b; Samuelson & Smith 1999; Soja, Carey & Spelke1991; Markman, 1991; Bauer & Mandler, 1989.) Yet the mechanisms responsible for this systematicity in performance are not yet specified. The research reported in this paper seeks insights into these mechanisms and is motivated by both contemporary and traditional ideas about the nature of categories. The first idea is the more modern one of emergent categories that reflects momentary goals as well as long term knowledge. The second idea is the more traditional one of stable categories that reflects regularities in experience. Considered jointly these two ideas suggest a new way of thinking about how language influences children's category formation and fosters cultural differences.

Contemporary Idea

Barsalou (1983) introduced the idea of an ad hoc category, a category formed in the moment in the service of solving a specific problem, for example one might from a category of "all the objects on my desk I can use to draw a straight line" or "all the things I can eat in this restaurant that I can afford." Our ability to form coherent ad hoc categories clearly depends on long-term knowledge of objects, their properties, and goals. The idea, however, is that specific bits of knowledge and the current input are melded to form a category adaptively fit to the moment. "Ad hoc categories should come to mind only when primed by current goals. Laurence W. Barsalou, 1983, (p. 223)"

Traditional Idea

The traditional idea of categories contrasts with the on-line flexibility of Barsalou's ad hoc categories. Traditionally, categories have been seen as fixed, mental structures that one either has or does not. It is this traditional idea about categories that make Whorf's (1956) claims about the role of language in categorization so contentious. Whorf argued that concepts are the product of language and so speakers of different languages "have" different concepts. But viewed through a more modern lens, Whorf's ideas present another possibility. "… all observers are not led by the same physical evidence to the same picture of the universe, unless their linguistic backgrounds are similar."(1956:214)—(p. 6 Gumperz & Levinson, 1996) Here is the idea: Language operates both as a contextual and long-term force an on-line category formation, causing speakers of different languages to attend to different properties in the task and thus to create different ad hoc categories. We test this idea in a study of English- and Japanese-speaking children's attention to cues distinguishing animates and inanimates.

Animates and inanimates

All languages distinguish between animates and inanimates and do so in a variety of ways. However, English and Japanese differ in their systems for individuating objects, and specifically in whether they treat animates as special discrete kinds. English treats both animate and inanimate objects as discrete and countable, that is, names for animate and inanimate objects are count nouns that take the plural. Substances,

in contrast, are named by mass nouns and not pluralized but instead take continuous quantifiers. In contrast, Japanese treats only animates as discrete. Nouns in Japanese are not pluralized, but names for animates can take an optional plural form. Further, the classifier system used for counting distinguishes animates from inanimate objects, but does not distinguish substances from inanimate objects. If Whorf is right, these linguistic differences might lead Japanese speakers more than English speakers to attend to properties relevant to animal categories. If instead Barsalou is right about the online creation of categories, cues suggesting the relevance of animacy properties should make speakers of both languages attend to properties relevant for animal categories. Finally, if both Whorf and Barsalou are right, speakers of both languages may be differentially sensitive to these cues and thus form different categories online.

Task

One task commonly used to examine children's category formation is the novel noun generalization task (e.g., Landau, Smith & Jones, 1988.) In this task, children are presented with a novel object, told its name, and then asked what other objects have the same name. Since both the objects and names are novel, this task measures category creation. Research has shown that children smartly use information about the objects and information from linguistics in the category formation (See, Smith, 1995, for a review). Critical to the present research are well replicated findings showing that when the named novel object has the properties of an artifact (solid, angular, complex shapes), children from categories based on shape but when the named novel object has properties of animals (e.g., eyes), children form categories based on a joint similarity in shape and texture (Jones et. al, 1991). We use this task and children's creation of shape-based versus shape plus texture based categories as a measure of the role of language history and on-line linguistic cues in children's category formation. More specifically, the experiments examine the interaction of the immediate linguistic cues, the individual's history of using those linguistic cues, and perceptual properties of the to-be-classified objects.

Experiment 1

The first experiment examined Japanese-speaking children's use of both linguistic cues and perceptual cues in category formation. The design crosses two levels of linguistic cues, one clearly marking the object as animate, one clearly marking the object as inanimate—with three different perceptual cues, cues that strongly suggest an animate thing, cues that weakly suggest an animate thing and cues that strongly suggest an inanimate thing.

The linguistic cues are *aru* and *iru* . In locative constructions (e.g., There is a cup) *aru* is obligatorily used with inanimate objects and *iru* is obligatorily used with animate objects. This is a highly salient and pervasive lexical contrast in Japanese. Figure 1 illustrates the 3 kinds of perceptual cues: rounded objects with eyes, rounded objects with appendages that might be viewed as legs, and angular objects with no properties suggestive of animate things. We ask: How do Japanese-speaking children combine these perceptual and linguistic cues when forming a new category?

Method

Participants Sixty monolingual Japanese-speaking children who were 21.06 to 45.7 months old participated. Participants' mean age was 34.51 months. All the participants were recruited from Niigata, Japan. All children participated with their parents.

Stimuli, materials and procedure Children participated in one of the six conditions that resulted from crossing the two linguistic cues (*iru/aru*) with the three levels of perceptual cues. In each condition children were tested in two blocks of 12 trials. In each block, a unique exemplar was named with a novel name and the child was asked whether that name also labeled each of 6 test objects. These six test objects were each queried twice in a randomly determined order. Three of these were control objects designed so that children should respond the same way regardless of whether they perceived the named object as depicting an animate or inanimate. One control object matched the exemplar in all features thus all children should say "yes" the name of the exemplar applies to this object. Two control objects differed from the exemplar in both shape and texture (or shape and color), thus all children should say "no" the name of the exemplar does not apply to these objects. The three diagnostic test objects matched the exemplar in shape and texture, shape and color, or only shape. If children perceive the exemplar as an animate, they should say the name applies only to the shape-and –texture matching object. If children perceive the exemplar as an artifact, they should say " yes" the name of the exemplar applies to all three diagnostic objects.

The sentence frames used in the experiment were presented in the following in the Iru and Aru conditions, respectively: "*Kokoni ____ ga iru yo*" and "*Kokoni ___ ga aru yo*." Test objects were queried as follows in the Iru and Aru conditions respectively: "*Kokoni ____ga iru kana*?" and "*Kokoni _____ ga aru kana*?"

The objects used in this experiment are, illustrated in Figure 1. The 3 control objects and 3 diagnostic test objects for the two test sets were constructed in the same way and are labeled in Figure 1 by the properties on which they match the named exemplar. All objects

were 3-dimensional, approximately 7cm x 7cm x 7cm, in size. The sample set illustrated in Figure 1 is stimuli with suggestive of animate cues: all objects in the pre-training set, the keppuru set, and tema set had appendages made of pipe cleaners.

Results

When the exemplar had eyes, a clear cue indicating the depiction of an animate, children generalized the name to new instances that matched in both shape and texture. The linguistic cues of *iru* and *aru* had no effect on performance. When the exemplar was angularly shaped and presented no cues suggesting an animate thing, children generalized the name to all test objects matching in shape. The linguistic cues of *iru* and *aru* had no effect on performance. However, when the exemplar presented weak perceptual cues suggesting an animate thing, the linguistic cues had a dramatic effect. The children generalized the name in the context of *iru* (the animate form) only to test objects that matched the exemplar in shape and texture just as they did when the exemplar had eyes. However, in the context of *aru* (the inanimate form) , the children generalized the name to all objects that matched the exemplar in shape just as they did when the exemplar was clearly an artifact. These conclusions were confirmed by analyses of "yes" responses (the name applies) on the diagnostic trials.

Children's proportions of "yes" responses on these trials were submitted to analysis of variance for a 2 (Linguistic cues) X 3(perceptual cues) mixed design. The analysis revealed the main effect of Linguistic cues, $F\ (1,54) = 15.834$, $p < .001$, the main effect of perceptual cues, $F\ (2,54) = 14.132$, and the interaction between perceptual cues and linguistic cues, $F\ (2,54) = 9.073$, $p < .001$. The "yes" responses for these diagnostic test objects in the 6 conditions are shown in Figure 2. These results show: (1) clear effects of perceptual cues on category formation. (2) clear effects of linguistic cues, and (3) the dominance of perceptual over linguistic cues, at least for these children in this task context.

Experiment 2

Experiment 1 showed Japanese-speaking children's sensitivity to linguistic cues in category formation, when the perceptual cues were not strongly pointing in some other direction. The goal of Experiment 2 was to replicate this finding with English-speaking children. Specifically, children were provided with stimulus objects that presented weak cues suggestive of animacy, appendages that could be seen as legs. The exemplar was named in one of two sentence frames that used different verbs; one using a verb suggesting an animate kind, the other using a neutral verb. The verbs used were "wants" and "goes": The exemplar was named saying, "*This mobit wants to stay here."* or *"This mobit goes here"*

Method

Participants Twenty monolingual English-speaking children between the ages of 23 to 43 months participated. The mean age was 31.7 months. The experimental sessions were held in Bloomington, IN. Participation of children was voluntary.

Stimuli, materials, design and procedure All aspects of the stimuli, procedure and design were identical to Experiment 1 with the exceptions of the type of stimuli and the sentence frames in which the novel names were presented. We used only one perceptual level for stimuli: ambiguous objects (see Figure. 1). The sentence frames used in English were non-locative constructions that had plausible animate/inanimate distinction "wants" for animate and "goes" for inanimate or neutral. The novel words employed to name the exemplars in the Experiment 2 were slightly altered to sound natural in English (e.g., *mobito*/mobit; keppuru/kipple; tema/teema).

Results

As shown in Figure 3, the English-speaking children generalized the names for these objects in the same way in both linguistic conditions, to all test objects that matched the exemplar in shape. This suggests children saw the objects as artifacts. The key result, however, is that they were unaffected by the linguistic cues. This contrasts with the Japanese-speaking children of Experiment 1 who categorized these same ambiguous objects differently in the two linguistic contexts.

Experiment 3

Are English-speaking children just less sensitive to linguistic cues? Perhaps "wants" is not as strong a cue suggesting animacy in English as "*iru* " is in Japanese. In Experiment 3, we used a sentence context containing personal pronouns (he/she) in an attempt to encourage children to form animal-like categories.

Method

Participants Twenty monolingual English-speaking children between the ages of 24 to 43 months participated. The mean age was 32.95 months. The experimental sessions were held in Bloomington, IN. Participation of children was voluntary.

Stimuli, materials, design and procedure All aspects of the stimuli, procedure and design were identical to Experiment 2 with the exception of the sentence frames in which the novel names were presented. The sentence frames used in English had either pronoun "she/he" or neutral subject "it" to refer the object

Results

English-speaking children showed clear (and reliable) effect of linguistic cue. As is evident in Figure 4, in the context of "he/she," children generalized the name to test objects matching the exemplar in shape and texture. In the context of "it," children generalized the name to new instances the same shape as the exemplar. Thus, we see clear on-line effects of linguistic cues on category formation in English-speaking as well as Japanese-speaking children.

The analysis revealed that the number of 'yes' responses to the shape matching test objects for the two groups of children differed reliably, $t=3.851$, $p<.005$. Given objects with features suggestive of animal limbs, English-speaking children provided with pronoun *"she/he*" were significantly more likely to form a narrower category based on both shape and texture than children with the neutral subject "*it*" only new instances that matched in shape alone.

Clearly, these pronouns activated different conceptual understanding that directed children's attention to different properties.

Experiment 4

The evidence thus far fits the contemporary vision of systematic categories created on-line and in-the-moment from the combination of perceptual cues and linguistic cues. But what of the long-term effects of learning particular language with a particular structure on on-line category formation. Although both English and Japanese have many linguistic devices and contrasts organized around animacy, the Japanese language through its system of individuation is arguably more concerned with animacy than English. Are linguistic cues suggesting animacy thus stronger for Japanese-speaking children than English-speaking children? To test this, we presented both English- and Japanese-speaking children with a novel angular artifact with no cues even remotely suggestive of animate thing. We named the object with a novel name in a linguistic context loaded with multiple cues indicating the conceptualization of the object as an animate.: for English *"See, who do you think this is? He is a Mobit! Isn't this mobit cute? There might be some more mobits that came to play with us!"* and for Japanese, *"Hora, koko ni dare ga iru to omou? Kokoni iru nowa mobito kun. Mobito kuntte kawaii desyo? Hoka nimo motto ippai mobito ga asobini kiteirukamo shirenaiyo!*"

Method

Participants Ten monolingual English-speaking children who were 25 to 45 months old and 10 Japanese-speaking children who are 27.4 to 38 months old participated. English-speaking children's mean age at this study was 35.5 months, and Japanese-speaking children's mean age was 33.82 months. The English-speaking children were recruited from Bloomington, IN. The Japanese-speaking children were recruited from Niigata, Japan. All children participated with their volunteer parents.

Stimuli, materials, design and procedure All aspects of the stimuli, procedure and design were identical to Experiment 2 & 3 with the exceptions of the sentence frames in which the novel names were presented.

Results

As is evident in Figure 5, Japanese-speaking children generalized the name to all test objects that matched the exemplar in shape and texture, the pattern typical of animate things. In contrast, English-speaking children generalized the name to all instances like the exemplar in shape, the pattern typical of artifacts. The numbers of 'yes' response to the shape matching test objects for the two groups of children differ reliably, $t=7.577$, $p<.001$.

Apparently English-and Japanese-speaking children see these objects differently. English-speaking children form categories based on the perceptual cues, Japanese-speaking children follow the linguistic cues. It appears that different histories of using language lead children to make different use of on-line information, and thus, in-the-moment-of the task, form different categories.

General discussion

The Whorfian question is often conceptualized as asking whether speakers of different languages "have" different categories. This question does not make sense if categories are momentary creation. Certainly, children's categorizations in the novel noun generalization task used here are momentary creations, formed *de nouveau* from learning a single novel object named with a novel name. But both English- and Japanese-speaking 3 year olds readily and coherently create categories in this task using perceptual and linguistic information in the task as the basis for categorizing new instances. But how in-task, in-the-moment, information is used will also depend on the life time histories in using those cues in other contexts. In this way, we may see direct effects of the language one knows, not on the categories and concepts one has, but on the categories and concepts one creates to solve a specific task.

Figures

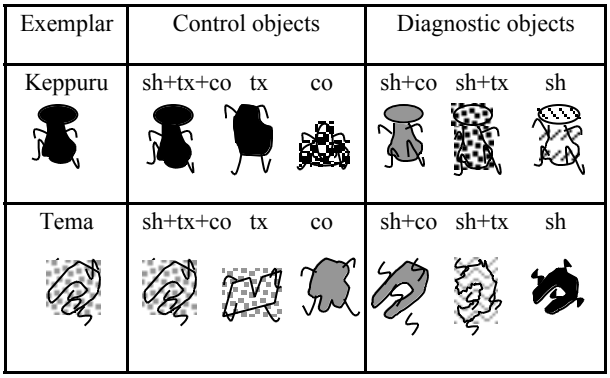


Figure 1: Test set stimuli with suggestive animacy features.

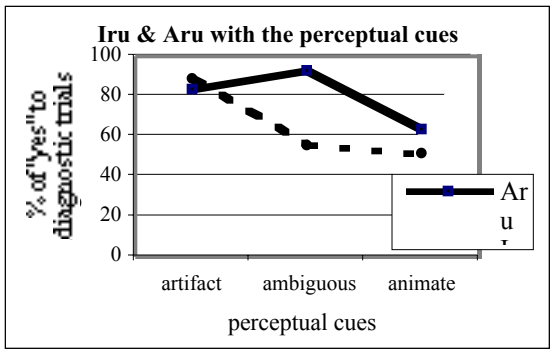


Figure 2: Proportion of "yes" responses by Japanese-speaking children with two different linguistic cues to test objects that matched the exemplar by shape. On the x-axis, objects are labeled by three perceptual cues.

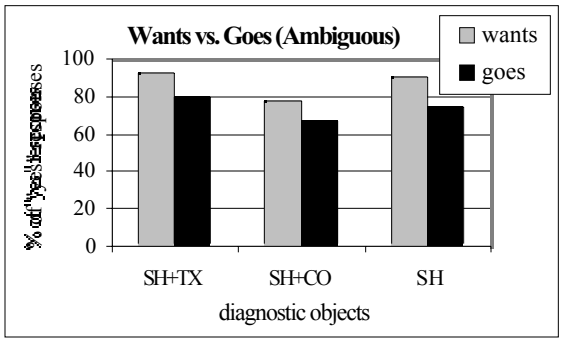


Figure 3. Proportion of "yes" responses by English-speaking children with two different linguistic cues to test objects. On the x-axis, objects are labeled by the properties matched to the exemplar.

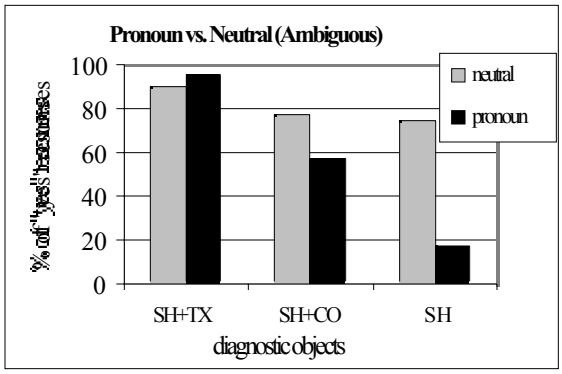


Figure 4. Proportion of "yes" responses by English-speaking children with two different linguistic cues to test objects. On the x-axis, objects are labeled by the properties matched to the exemplar.

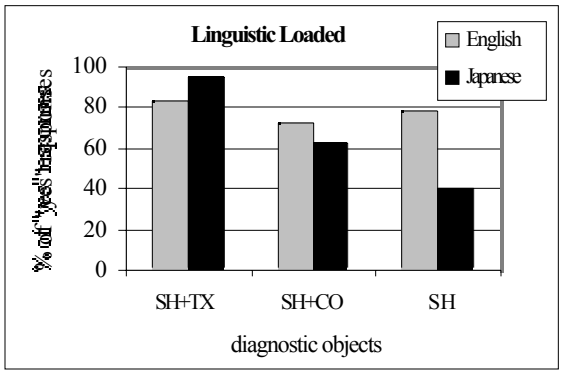


Figure 5. Proportion of "yes" responses by Japanese- and English-speaking children with a maximum linguistic cues to test objects that are clearly artifacts. On the x-axis, objects are labeled by the properties matched to the exemplar.

References

Barsalou, L. W. (1983). Ad Hoc categories. *Memory and Cognition*, *3*, 211-227

Bauer, P.J. & Mandler, J.M. (1989) Taxonomies and triads: Conceptual organization in one- to two-year olds. *Cognitive Psychology, 21*, 156-184.

Gumperz, J. J., & Levinson, S. C. (1996) Introduction; Linguistic relativity.re-examined. In J. J. Gumperz & S. C. Levinson (Eds.), Rethinking linguistic relativity (pp-1-18). Cambridge: Cambridge University Press

Jones, S., Smith, L., & Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child Development, 62,* 499-516.

Landau, K. B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3*, 299-321.

Landau, B., Smith, L. B., & Jones, S. (1998a) Object shape, object function, and object name. Journal of *Memory and Language, 38*, 1-27.

Landau, B., Smith, L. & Jones, S. (1998b) Object perception and object naming in early development. *Trends in cognitive science, 2*, 19-24.

Mandler, J., M., Bauer, P. J., and McDonough, L. (1991) Separating the sheep from the goats; Differentiating global categories. *Cognitive Development*, *3*, 247-264

Markman, E. M. (1991). The whole object, taxonomic, and mutual exclusivity assumptions as initial constraints on word meanings. In J. P. Byrnes & S. A. Gelman (Eds.), Perspectives on language and cognition: Interrelations in development (pp-72-106). Cambridge: Cambridge University Press

Samuelson, L. & Smith, L. B. (1999) Early noun vocabularies: Do ontology, category structure, and syntax correspond? *Cognition*. *73,* 1-33

Smith, L. B. (1995). Self-organizing processes in learning to learn words: Development is not induction. The Minnesota Symposia on Child Psychology. Volume 28. Basic and applied perspectives on learning, cognition, and development (pp. 1-32). Mahwah, NJ: Lawrence Erlbaum Associates.

Soja, N., Carey, S., & Spelke, E. (1991). Ontological categories guide young children's inductions of word meanings: Object terms and substance terms. *Cognition, 38,* 179-211.

Whorf, B. (1956). Language, thought and reality: Selected writings of Benjamin Lee Whorf (J.B. Carroll, Ed.). Cambridge, MA: MIT Press.

Base-Rate Neglect in Pigeons: Implications for Memory Mechanisms

Thomas R. Zentall (zentall@pop.uky.edu)
Tricia S. Clement (tcharb1@pop.uky.edu)
Department of Psychology, University of Kentucky
Lexington, KY 40506 USA

Abstract

In delayed matching-to-sample, there is an initial (or sample) stimulus and two test (or comparison) stimuli. When pigeons are trained to match, they presumably choose between the two comparison stimuli according to their ability to remember the sample. When the sample cannot be remembered, comparison choice should depend on the history of reinforcement associated with each of the comparison stimuli (i.e., the comparison base rates). In the present research, pigeons acquired two matching tasks in which samples S1 and S2 were each associated with one of two comparisons C1 and C2 (equal experience with both trial types), and samples S1 and S3 were each associated with one of two other comparisons C3 and C4 (equal experience with both trial types). As the retention interval increased, the pigeons showed a bias to choose the comparison (C1 or C3) associated with the more frequently occurring sample (S1). Thus, pigeons are sensitive, not just to the probability of reinforcement associated with the each of the comparison stimuli (i.e., the base rates) which were equal, but also to the (irrelevant) likelihood that each of the samples was presented (i.e., base-rate neglect).

Introduction

Humans are known to underestimate the effect of base rates associated with probability of being correct (Kahneman & Tversky 1972). In a classic problem proposed by Tversky and Kanemman (1980, p. 62), participants are told that 85% of the taxis in a city are green while only 15% are blue (the base rates). They are also told that a witness to a hit-and-run accident involving a taxi identified the taxi as blue. Furthermore, they are told that under similar conditions witnesses correctly identify the color of a taxi 80% of the time. When participants are then asked, "What is the probability that the taxi involved in the accident was actually blue?" most of them say that it is very likely that the taxi is blue. In making this judgement the participants fail to consider sufficiently the base-rate probabilities. When base rate is considered, the conditional probability of correctly identifying a blue taxi is p(blue|judgement correct) = p(blue) x p(correct) = .15 x .80 = .12, whereas the probability of saying it was blue when it actually was green is p(green|judgement incorrect) = p(green) x p(incorrect) = .85 x .20 = .17. This means that the probability of being correct under these conditions is only .12/(.12 + .17) = .41, or less than 50%. Thus, humans often fail to consider sufficiently the probability of being correct in the absence of the eye-witness information. Although there are certain conditions under which humans can be induced to perform more accurately (e.g., Gigerenzer & Hoffrage, 1995), base-rate neglect is likely responsible for many exaggerated fears such as air travel, walking the streets of New York City, and having one's children killed at school by a fellow student.

An analogous situation can be designed for an animal using a matching-to-sample task. Matching-to-sample is a conditional discrimination in which the identity of the initial or sample stimulus indicates which of two (or more) test or comparison stimuli is correct (Skinner, 1950). According to Hartl and Fantino (1996), comparison choice for pigeons should depend on two factors, the relative probabilities of reinforcement associated with the comparisons (i.e., the base rates) in the absence of the sample, and the conditional probability of each comparison being correct given presentation of one of the samples (i.e., the actual sample event or the evidence, given the base rates). In the case of matching-to-sample, the probability of reinforcement given the sample is typically 1.0. This ensures that the task has been adequately acquired and that the contingencies have been adequately experienced. Biases can be introduced by manipulating the ratio of samples and the probability of reinforcement for choices of the matching comparison (see Goodie & Fantino, 1995, for similar findings with humans, but see also Goodie & Fantino, 1996, for exceptions).

Control by the comparisons alone can be increased by degrading the samples at the time of comparison choice (i.e., by increasing the probability of poor memory, or in the taxi example, of an identification error). One way to degrade the samples is by introducing a delay between the offset of the sample and the onset of the comparisons. Assuming that the comparison stimuli are correct equally often over trials, and that the probability of reinforcement is the same for a correct response to each comparison, one would expect that with increasing delay, the slopes of the pigeons' retention functions would be quite similar (see Grant, 1991; White &

Wixted, 1999).

The analog to base rate in a matching task is the probability of being correct in the absence of information about the sample (i.e., the relative probability of reinforcement associated with each of the comparison stimuli). According to White and Wixted (1999), pigeons should be sensitive to base-rate probabilities, but generally the base rates and the probability of sample presentation are the same (both generally 0.5). In the present experiment we asked if pigeons are able to estimate the probability of a correct comparison response when the sample probabilities are different from the base rates. There are a number procedures that might be used to manipulate the relative frequency of sample (S) presentation while maintaining equal probability of reinforcement for comparison (C) choice (i.e., equal base rates). In the present experiment, we chose to introduce a second 2-sample-2-comparison matching task. Each of the two matching tasks involved a different pair of comparison stimuli but the two tasks shared a common sample. Thus, the two tasks can be represented S1-C1, S2-C2 and S1-C3, S3-C4 (with C1 and C2 always appearing together and C3 and C4 always appearing together). If each of the four trial types appears equally often, each of the comparisons would be associated with reinforcement on 25% of the reinforced trials. However, the same would not be true of the samples. S2 and S3 would each be presented on 25% of the trials, whereas S1 would be presented on 50% of the trials. Under conditions with no delay, one would expect a high level of matching accuracy and no bias. But if a delay is inserted between the offset of the sample and the onset of the comparisons, errors should increase. If comparison choice depends on the reinforcement contingencies associated with comparison choice, errors should not result in a comparison bias. In the absence of memory for the sample, the probability of reinforcement of comparison choice should be 50% for either comparison in either task. Furthermore, if there is memory for the sample, the conditional probability of reinforcement associated with comparison choice should be the same for either comparison in either matching task. However, if pigeons show a bias by using their reference memory of sample presentations, they should access more instances of S1 than of either S2 or S3 and a bias to choose C1 and C3 may result.

Method

Subjects

The subjects were eight White Carneaux pigeons, purchased as retired breeders (5-8 years old) from the Palmetto Pigeon Plant (Sumter, SC). The pigeons were maintained at 80% of their free-feeding body weights throughout the experiment and were caged individually with grit and water continually available in the home cage. The pigeons were maintained on a 12:12-h, light:dark cycle. All pigeons had previously served in an unrelated study involving simple simultaneous discriminations.

Apparatus

The experiment was conducted in a standard BRS/LVE (Laurel, MD) sound attenuating pigeon test chamber. Three rectangular response keys (2.5 cm high x 3 cm wide and 1 cm apart) were aligned horizontally and centered on the response panel. Mounted behind each response key was a 12-stimulus inline projector (Industrial Electronics Engineering, Series 10, Van Nuys, CA,) that could project a red hue or a green hue onto the any of the three response keys or a plain white field onto the center response key. In addition, the left and right projectors could project a white circle and a white dot. A houselight located at the center of the chamber ceiling provided general illumination. A rear-mounted grain feeder was centered horizontally on the response panel midway between the pecking keys and the floor of the chamber. When operated, the feeder was accessible through a 5.0 x 5.5 cm lit aperture in the response panel. Reinforcement consisted of 2.0-s access to Purina Pro Grains. White noise and an exhaust fan mounted on the outside of the chamber masked extraneous noise. The experiment was controlled by a microcomputer located in an adjacent room.

Procedure

Training All pigeons were placed directly on 0-s-delay matching-to-sample training. At the beginning of each trial, the center key (sample) was illuminated. Following 10 responses to the sample, the sample was turned off and the side (comparison) keys were illuminated. Comparison stimuli were presented randomly with respect to location, with the restriction that a particular hue could not occur on the same side key for more than three consecutive trials. One response to either comparison constituted a choice and terminated the trial. Correct comparison responses resulted in a 2-sec presentation of food and a 10-sec intertrial interval. Incorrect choices resulted in the 10-sec intertrial interval alone.

For each pigeon, training consisted of a hybrid matching task involving three sample stimuli (one per trial) and two pairs of comparison stimuli (one pair on each trial). On one fourth of the trials, one of the hues served as the sample (S1) with red and green comparison stimuli (C1 and C2) on the side keys and, for example, red was correct. On another fourth of the trials, a different hue sample (S2) was presented with the red and green comparison stimuli and, for example, green was correct.

On half of the remaining trials, S1 was again presented as the sample and circle and dot were presented as the

comparisons (with, for example, dot correct). On the remaining fourth of the trials a third hue was presented as the sample (S3) and circle and dot were presented as the comparisons (with circle correct).

The three sample hues were counterbalanced such that each hue served as the one-to-many sample for 2-3 pigeons and each of the remaining samples was associated with the hue comparisons for at least one pigeon. Sessions consisted of 96 trials and were conducted 6 days a week. For each pigeon, criterion was met when the correct comparison for each trial type was chosen on at least 90% of those trials for two consecutive sessions. Following criterion performance, each pigeon received five sessions of overtraining.

Retention test On the following session, each pigeon was transferred to a mixed-delay matching procedure in which the offset of the sample was separated from the onset of the side keys by a dark retention interval of 0, 2, 4, or 8 s. For each of the trial types, there was an equal number of trials involving each retention interval. The retention test consisted of 2 sessions and the reinforcement contingencies were the same as they were during training. In all analyses of results, the .05 level of statistical significance was adopted.

Results

Training

Sessions to criterion (two successive sessions at 90% correct) for the one-to-one portion of the task was 10.2 when the comparisons were hues and 11.1 when the comparisons were shapes. Sessions to criterion for the one-to-many portion of the task was 13.6 when the comparisons were hues and 13.8 when the comparisons were shapes. A mixed-effect analysis of variance performed on the acquisition scores, with task (one-to-one vs. one-to-many) and comparison dimension (hues vs. shapes) as factors, indicated that neither effect nor the interaction was statistically reliable, $F(1,7) = 2.08$, >1, and >1, respectively.

Retention Test

Data from the retention test were pooled over the 2 test sessions and were subjected to a repeated-measures ANOVA, with task component (one-to-one vs. one-to-many) and Delay (0, 2, 4, and 8 sec) as factors. Most critically, the ANOVA indicted that there was a significant Task Component x Delay interaction, $F(3,21) = 4.37$. There was also a significant effect of Delay, $F(3,21)$ 44.01. The effect of Task Component was not quite significant, $F(1,7) = 4.79$. The retention data are presented in Figure 1.

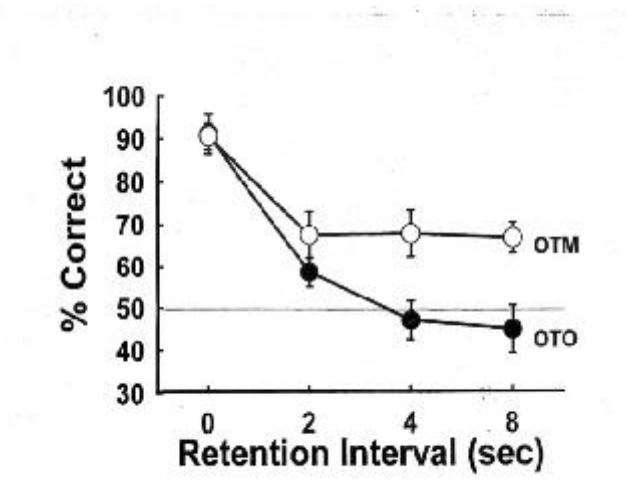


Figure 1. Retention functions following training in which two and samples, S1 and S2, were associated with comparison stimuli, C1 C2, respectively and S1 and S3 were associated with comparisons C3 and C4, respectively. Thus, S2 and S3 were involved in one-to-one matching (OTO) with C2 and C4, while the third sample, S1, was associated with two comparison stimuli, C1 and C3 (one-to-many matching, OTM). In training and test, each comparison was associated with reinforcement on 50% of the trials and C1 and C2 always appeared together as did C3 and C4.

Discussion

According to traditional instrumental views of conditional discrimination learning (i.e., Hartl & Fantino, 1996), the probability of a comparison choice should be determined by the conditional probability associated with each comparison stimulus, given the sample, and, if the sample is unavailable or forgotten, with the probability of reinforcement associated with each comparison (independently of the sample). Thus, the choice a particular comparison (e.g., C1) should depend on both the number of sample-comparison pairings (e.g., S1-C1) that are followed by reinforcement, as well as the number of reinforcements associated with that comparison, independent of the sample (Wixted, 1993). In the present experiment, the conditional probability of reinforcement associated with each of the comparisons, given one of the samples, was equal. Furthermore, the probability of reinforcement associated with choice of either comparison was also equal. Thus, in the present experiment, given presentation of C1 and C2, the only relevant sample-comparison associations determining comparison choice should be S1-C1 and S2-C2. If so, delay-induced sample degradation should have had a symmetrical effect on comparison choice and the retention functions should have been parallel and overlapping.

In the present experiment, clearly divergent retention functions were found. These results require the modification of current theories of delayed conditional discrimination performance (e.g., White & Wixted, 1999) because pigeons choice behavior is influenced not only by the probability of reinforcement associated with responding to each of the comparison stimuli and to the conditional probabilities associated with choice of the comparison stimuli as a function of memory for the sample but also by the relative frequencies of the samples. When delays are introduced, as the delay increases, pigeons have an increasing tendency to select the comparison associated with the more frequently presented sample, even though that sample was not presented more often than the alternative sample in the context of either comparison pair. It is as if, on trials when memory for the sample is poor, presentation of the comparisons causes the pigeons to consult their reference memory for the overall probability of sample presentation (independent of the comparison pair).

Of broader interest, such use of reference memory in delayed matching may be a general phenomenon. However, the use of sample frequency independently of other more relevant measures may be apparent only with a design such as that used in the present research because in the more typical design, either hypothesis makes the same prediction.

Alternatively, in the present experiment, although the pigeons had equal opportunity to acquire each of the four sample-comparison associations, the more frequent presentations of the S1 sample could have allowed it to be more efficiently coded, better maintained in memory, or more easily retrieved from memory. That is, at the time of comparison choice, when the S1 stimulus had been the sample, it may have been more accessible than the S2 or S3 stimuli were when they had been the sample. But if the difference in slope of the retention functions was attributable to differences in sample accessibility at the time the comparisons were presented, both the S1 and the S2/S3 functions should have approached 50% correct with increasing retention interval. Instead, the S1 retention function appears to have leveled off, while the S2/S3 retention function declines below chance at delays of 4 and 8 sec. Such retention functions suggest that rather than better retrieval of the S1 sample, the pigeons developed a comparison bias to choose the comparison associated with the more frequently presented sample.

This comparison bias in pigeons is analogous to the base-rate neglect shown by humans when they fail to consider sufficiently the base-rate probability of occurrence of an event.

References

Gigerenzer, G. & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. Psychological Review, 102, 684-704.

Goodie, A. S., & Fantino, E. (1995). An experimentally derived base-rate error in humans. Psychological Science, 6, 101-106.

Goodie, A. S., & Fantino, E. (1996). Learning to commit or avoid the base-rate error. Nature, 380, 247-249.

Grant, D. S. (1991). Symmetrical and asymmetrical coding of food and no-food samples in delayed matching in pigeons. Journal of Experimental Psychology: Animal Behavior Processes, 17, 186-193.

Hartl, J. A., & Fantino, E. (1996). Choice as a function of reinforcement ratios in delayed matching to sample. Journal of the Experimental Analysis of Behavior, 66, 11-27.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgement of representativeness. Cognitive Psychology, 3, 430-453.

Skinner B. F. (1950). Are theories of learning necessary? Psychological Review, 57, 193-216.

Tversky, A, & Kahneman, D. (1980). Causal schemas in judgements under uncertainty. In M. Fishbein (Ed.), Progress in social psychology (pp. 49-72). Hillsdale, NJ: Erlbaum.

White, K. G., & Wixted, J. T. (1999). Psychophysics of remembering. Journal of the Experimental Analysis of Behavior, 71, 91-113.

Wixted, J. T. (1993). A signal detection analysis of memory for nonoccurrence in pigeons. Journal of Experimental Psychology: Animal Behavior Processes, 19, 400-411.

Author Notes

This research was supported by National Institute of Mental Health Grants 55118 and 59194.

Correspondence should be addressed to Thomas R. Zentall, Department of Psychology, University of Kentucky, Lexington, KY 40506-0044. Email, zentall@pop.uky.edu

Member
Abstracts

Explanations of words and natural contexts: An experiment with children's limericks

Greg Aist (aist@cs.cmu.edu)
Project LISTEN, Carnegie Mellon University, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Introduction

Project LISTEN's Reading Tutor listens to children read aloud, and helps them learn to read. Here, we used the Reading Tutor to study whether adding child-friendly definitions to natural text would help children learn new words. We compared four conditions:

1. No encounter
2. In a definition alone
3. In a story alone
4. In a story and a definition

This study took place at a July 2000 reading and math clinic at a low-income urban elementary school in Pittsburgh. Each student was scheduled to spend 30 minutes per weekday on the Reading Tutor.

Experiment design

We used eight children's limericks by Edward Lear (19th c.), with one target word each: *dolorous*, *laconic*, *imprudent*, *innocuous*, *mendacious*, *oracular*, *irascible*, or *vexatious*. The target word was always the second word in the last line. Our text selection controlled for genre, author, intended audience, (approximate) word frequency, part of speech, and general semantic class:

> There was an Old Man of Cape Horn,
> Who wished he had never been born;
> So he sat on a chair,
> Till he died of despair,
> That dolorous Man of Cape Horn.

We wrote target word definitions in a consistent style using ordinary language, following McKeown (1993). For example: "We can say someone is dolorous if they are mournful, or feel really bad." To reduce variance from first- or last-item effects, we held constant the order of presentation of the limericks. Each student saw two target words per condition. Word-to-condition assignment was set for each Reading Tutor computer.

One or two days later – depending on attendance – we gave each student a paper questionnaire with two items per word. "Have you ever seen the word *dolorous* before?" tested familiarity, and "If someone is dolorous they must be… angry; sad; tired; afraid." tested word knowledge. To exclude memorization, the definitions and the test answers used different words.

In all, 29 students who had just finished 2nd - 5th grades completed the experiment, for a total of 232 trials, 58 trials for each of 4 conditions.

Results and discussion

To explore the effects of explanations and limericks, we used logistic regression – modeling a binary outcome variable using several categorical factors as input. If a factor's coefficient was significantly greater than zero, than that factor affected the outcome variable.

Word familiarity. Seeing an explanation helped, at $p < 0.001$: coefficient $1.08 \pm .32$; 99.9% confidence interval (CI) .02, 2.15. Seeing the word in a limerick showed only a weak trend, at $.50 \pm .32$; 90% CI -.02, 1.03.

Word knowledge. The results for word knowledge were more nuanced: a (not significant) trend for explanations ($.24 \pm .31$), but none for limericks ($-.05 \pm .31$).

However, younger students were about at chance:

2nd grade, 19/72 right (26%) 3rd grade, 18/72 right (25%)
4th grade, 16/56 right (29%) 5th grade, 10/32 right (31%)

For 4th and 5th graders, however, in a main-effects-only model, explanations helped ($p < .10$): $.89 \pm .52$, with 90% CI .04, 1.74. There was not an effect for limericks, at $-.13 \pm .51$. (Disaggregation by grade is exploratory.)

Conclusions. Thus in terms of learning word meaning, only explanations seemed to help – and only for fourth and fifth graders. These effects are neither same-day recency nor simple memorization. Aist (Aist 2000 ch. 6) discusses further.

Acknowledgments

This paper is based on work supported in part by the National Science Foundation under Grant Nos. REC-9720348 and REC-9979894, and by the first author's Harvey and NSF Graduate Fellowships. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation or the official policies, either expressed or implied, of the sponsors or of the United States Government. Dr. Jack Mostow directs Project LISTEN; http://www.cs.cmu.edu/~listen lists other team members.

References

Aist, G. 2000. Helping Children Learn Vocabulary during Computer-Assisted Oral Reading. Ph.D. dissertation, Language Technologies Institute, CMU.

McKeown, M. G. 1993. Creating effective definitions for young word learners. *Reading Research Quarterly* 28(1), 17-31.

Understanding death as the cessation of intentional action: A cross-cultural developmental study

H. Clark Barrett (barrett@mpib-berlin.mpg.de)
Center for Adaptive Behavior and Cognition
Max Planck Institute for Human Development
Lentzeallee 94, Berlin, Germany

Introduction

Taken together, the developmental literatures on children's understanding of intentional action and on children's understanding of death present a sort of paradox: while the former literature shows that even very young children have an intuitive grasp of the goal-directed nature of behavior that characterizes animate, living things, the latter seems to show that children do not realize until a much later age that the capacity to act intentionally ceases irreversibly when an organism dies. From an adaptationist perspective, this is perhaps surprising, given the potential adaptive value of being able to judge the capacity or incapacity of an animal to act, and perhaps to do harm. In many states which are characterized merely by a lack of motion – sleep, for example – an animal may still wake up and attack. When dead, however, it cannot. It may thus be adaptive for young children to understand that death, as opposed to sleep and other states of temporary inaction, entails the permanent cessation of the capacity for intentional action. And, given that an intuitive grasp of the capacity for intentional action in animate living things exists at a very young age, it should be available to serve as a conceptual substrate for an early understanding of death.

Study design

The present study was designed to probe for an understanding of death as the cessation of the capacity for intentional action by comparing judgments of 3 to 5 year old children about hypothetical dead and sleeping animals (human and nonhuman). Children were first asked questions about an awake animal, then about the same animal either asleep or dead. The target questions required children to judge whether the sleeping or dead animal would be capable of movement in general or in response to being touched, awareness of an external stimulus (someone moving nearby), and / or of having an emotion. The intentional action theory of death understanding predicts that children will exhibit the clearest understanding of death in response to questions involving movement and movement in response to stimulus, as opposed to questions that do not involve intentional action (e.g., awareness only). There was no predicted lower bound on the age of emergence of reasoning abilities probed here.

The study was conducted in two parts, both using the same interview procedures and questions. The first part of the study was conducted with 70 preschool and kindergarten children, age 3 to 5 years, in Berlin, Germany. The second part of the study was conducted with 70 Shuar children, age 3 to 5 years, in six small rural villages in the Amazon region of Ecuador. Because these populations vary widely both in exposure to various kinds of cultural and environmental input (e.g., television and films, direct contact with animals, personal experience with death), they were selected for comparison to determine whether the development of the kind of understanding of death probed here depends significantly on cultural inputs such as television, and / or aspects of personal experience such as contact with animals or firsthand experience with death. Parents in both populations were surveyed about relevant experiences of their children which might have influenced understanding of death, such as religious background, exposure to representations of death on television, and personal experience with death of animals or people.

Results

Both Shuar and German children demonstrated an understanding of death as the cessation of the capacity for intentional action by the age of 4. By this age, the large majority of children clearly distinguish sleep from death in this regard. In addition, many of these children understand that the capacity for subjective experience independent of action ceases in death as well, though performance on these questions, as predicted, was not as high at this early age as performance on questions involving intentional action.

The results of the study show that by age 4, children understand at least one aspect of death which is crucially important from an adaptive perspective: dead things can no longer act. Although this result stands in contrast to much of the developmental literature on death understanding, which suggest that children's understanding of death at this age is poor, it is consistent with a view of cognitive development which holds that development has been shaped by a history of selection for adaptive reasoning and decision making abilities.

Working Memory Processes During Abductive Reasoning

Martin Baumann (martin.baumann@phil.tu-chemnitz.de)
Josef F. Krems (krems@phil.tu-chemnitz.de)
Department of Psychology, Wilhelm-Raabe-Str. 43
Chemnitz University of Technology
09107 Chemnitz, Germany

Introduction

Abductive reasoning is the process of finding a best explanation for a given set of observations. It is an essential feature of many real world tasks like medical diagnosis, discourse comprehension, and scientific discovery. Such problems often need the processing of an amount of information far beyond the capacity limits of working memory (WM). But on the other hand, working memory is expected to play a central role in human reasoning. On the basis of a computational model of abductive reasoning (Johnson & Krems, 2000) and of theories of text comprehension we propose a mechanism that reduces WM load during abductive reasoning. It suggests that only unexplained symptoms are kept in working memory with explained symptoms are transferred to long-term memory reducing WM load.

From this model it follows that unexplained observations should be more available in a recognition or recall task during abductive reasoning than explained ones. We tested this prediction in three experiments each using a different memory task to test the availability of observations.

Experimental Studies

The Experimental Task

In all experiments a task (BBX) was used where participants had to discover the hidden state of a system through indirect observations. The observations were presented sequentially to the participants. Only the current observation was visible. In each trial, after a variable amount of observations, the participants had to perform a memory task testing the availability of a given observation. The major manipulation in all experiments was whether this observation was already explained at the time of the memory task or not. That is, whether the participant had received the neccessary additional information to explain the observation and actually generated a hypothesis explaining this observation.

Results and Discussion

In the first experiment we used a recognition test as memory task to test the availability of the relevant observation. In the second experiment the recognition test was replaced with an implicit memory task. The mental availability of explained and unexplained observations were tested here by presenting a probe hypothesis that had to be judged with regard to its compatibility with observations presented until then.

The results of the first experiment showed that unexplained observations are recognised significantly faster than explained ones, consistent with model predictions. Regarding the recognition accuracy there was no significant effect of interval or explanation status. We also found that maintaining an unexplained observation in WM slows down the recognition and reduces the recognition accuracy for other observations.

The second experiment showed contrary to the model's predictions a tendency of explained observations being forgotten more often with increasing number of intervening observations than unexplained ones. This result suggested that observations are held actively in WM until they are explained. After an explanation was generated they are lost from WM. The result also indicates that explained observations are not integrated in a representation in long-term memory. This interpretation was confirmed in a third experiment showing that participants memory for explained and unexplained observations in an unexpected recall test after the interruption of the reasoning task was equally low.

General Discussion

The results confirmed the hypothesis that unexplained observations are actively hold in WM during abductive reasoning until a causal explanation can be generated. Contrary to the predictions of the model these explained observations seem not to become integrated into a representation in long-term memory, but are simply forgotten. But this could be due to the structure of the reasoning task we used, which makes the construction of an integrated representation rather difficult. Therefore in future investigations we need to use a task providing a richer structure, more comparable to real world tasks like medical diagnosis.

References

Johnson, T.R., & Krems, J.F. (2000). *Use of Current Explanations in Multicausal Abductive Reasoning.* Manuscript submitted for publication.

Organizing Features into Attribute Values

Dorrit Billman (dorrit.billman@psych.gatech.edu)
Carl Blunt (gte138r@prism.gatech.edu)
Jeff Lindsay (gte457e@prism..gatech.edu)
School of Psychology, Georgia Institute of Technology
Atlanta, GA 30332 USA

The problem of representation change is important for understanding and developing both natural and artificial intelligence. People form new chunks or perceptual units from experience (Schyns & Rodet, 1997) and build up new, continuous, dimensions, (Goldstone, Lippa, and Shiffrin, 2001).

We investigate representation change reorganizing unrelated features into alternative values of the same attribute. For example, 'fins', 'wings', and 'legs' might initially be unrelated properties but people might learn to reorganize them as values of a new attribute, LIMB. Representing properties as attribute values may be importantly different from representing properties as a collection of uncoordinated features, for interpreting novel properties and for projecting inferences.

We investigated perceptually-based attribute formation, in the context of learning about cell-like organisms. We use stimuli where an initial analysis into features is highly available, but there are alternative ways of organizing these features into attributes.

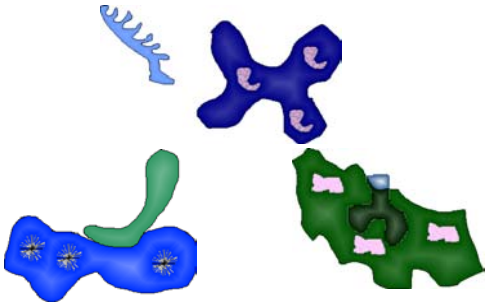


Figure1: Separate, Attached & Overlapping Units

Method

Stimuli were slides of cell-like organisms (*units)*, composed of a larger and a smaller element, each clearly bounded and differently colored. Units varied in the general shape of the larger element, its color, the small figures appearing internal to the larger element (e.g. its 'mitochondria'), and the shape of the smaller element. There were three values of each attribute.

We tried to influence whether participants saw the smaller element as parts of a larger organism versus as a buddy or other accompanying but distinct element, by spatial separation and by order of presentation. We predicted overlapping items would be most ambiguous and interpretation would be influenced by viewing order of Separate, Attached, and Overlapping blocks.

For initial exposure, subjects saw 3 blocks of six 1-unit displays (one block of each type) and described what they saw. Subjects then viewed 18 displays with one to five units ("slides with lower magnification") shown together and indicated how many organisms were on each slide, by circling and counting.

We varied the order subjects saw the Separate, Attached, and Overlapping blocks (Table 1). We expected the first exposure block to be influential. Does exposure order influence the final interpretation of the elements, as values of a new limb-like attribute or as values of a new buddy-like attribute?

Results & Discussion

Table 1 (col. 2) shows that displays had highest counts when subjects first saw the overlapping block, followed by the Attached and then Separated (item analysis $F(3,68)=181.9$). Interestingly, some subjects analyzed the small internal element as a separate organism, particularly in the Overlap First Conditions. Table 2 (col. 3) shows % of displays where subjects counted units all as 1 or all as 2 organisms; Overlap first subjects were less consistent ($F(3,48)=4.8$, $p<.01$)

Table 1

Conditions	Results	
Order of Blocks in Exposure Phase	AveCount /Unit	%consistent count strategy
1:Attach-Overlap-Sep (n=13)	1.75	78 %
2:Sep-Overlap-Attach (n=12)	1.88	81%
3: Overlap-Attach-Sep (n=9)	2.73	29%
4: Overlap-Sep-Attach (n=8)	1.88	60%

Presentation order influenced interpretation of elements as parts or separate organisms. The biggest influence of context came when the ambiguous displays were first. Further work will look at attribute change following concept learning.

Acknowledgements

Research supported by NSF grant 9732562 to Billman.

References

Goldstone, R.L., Lippa, Y., & Shiffrin, R.M. (2001). Altering object representations through category learning. *Cognition, 78,* 27-43.

Schyns, P.G. & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory and Cognition, 23(3)*, 631-696.

Attention Shift and Verb Labels in Event Memory

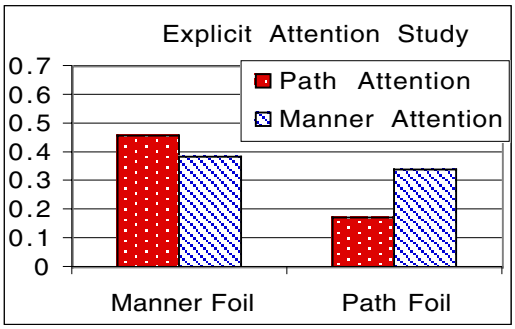
Dorrit Billman (dorrit.billman@psych.gatech.edu)
School of Psychology, Georgia Institute of Technology
Atlanta, GA 30332 USA
Michael Firment (mfirment@kennesaw.edu)
Psychology Department, Kennesaw State University
Kennesaw, GA 30144 USA

Events, unlike objects, pass by once and give only one opportunity to encode and retain information. How does the language accompanying events influence what is remembered? To what can people selectively attend, and what directs their attention?

In prior work (Billman & Krych, 1998) we found an interaction between type of visual recognition error and verb heard at encoding, such that people were more likely to notice changes in manner when they had heard an appropriate manner verb (e.g., hop, jog, skip) than an appropriate path verb (e.g., enter, cross, leave), and more likely to notice changes in path when they had heard a path verb than a manner verb. This effect was clearest when the foil events were changes such that the original verb would no longer appropriately apply to them. Thus, the effect of the encoding verb on memory might have resulted from using the verb as a discriminative cue during recognition, from using the verb to direct attention at encoding, or both.

Explicit Attention Study

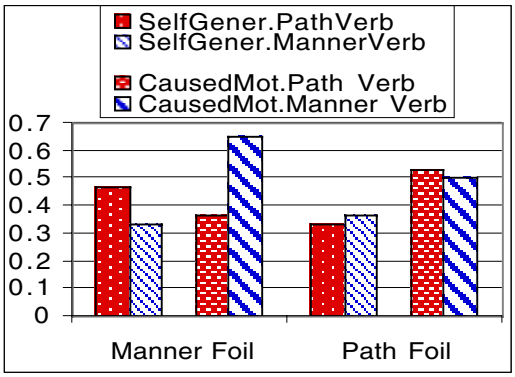
In the present studies we investigate whether attention can be directed to the manner or to the path of motion events. We used realistic, animated event clips of ordinary motion events. Half the events showed a self-generated motion (e.g. strolling man circles a car) and half a caused motion (e.g. cat rolls a pencil to make it



fall). Test items were very similar to the originals. In the Explicit Instruction Study we told participants to pay attention to path or to manner. As predicted, participants made fewer errors on items which changed the corresponding aspect of the event. (interaction of encoding condition x foil type, $F(2,32)=4.71$, $p<.05$). These data suggest a boundary for how much shifts of attention at encoding can aid recognition for our task and stimuli.

Verb Study

In the Verb Study, events were accompanied by a descriptive manner verb, path verb, or no language.



Verbs always described the moving figure (e.g. "rolling" or "falling"). If hearing a path or manner verb shifts attention this might benefit memory for the corresponding aspects, even when the verb itself provides no discriminative information at recognition. A three way interaction, $F(4,156) = 2.775$, $p < .05$, .among verb condition, type of event (self-motion or caused-motion), and type of recognition stimulus (path foil, manner foil, or original scene) on errors indicated the two types of events were affected differently. In the self-generated events, path verbs improved rejection of path foils and manner verbs improved the rejection of manner foils, as predicted, but the opposite pattern was found with the caused motion events. The mapping of the verb onto the moving figure is more complex in these events (it is the pencil not the animate entity which rolls) and this may be involved.

Acknowledgements

Research supported by NSF grant 9732562 to Billman.

Billman, D &Krych, M. (1998). Path and manner verbs in action: Effects of "Skipping" or "Exiting" on event memory. *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* Hillsdale, NJ: Lawrence Erlbaum Associates.

The Semantics of temporal prepositions: the case of IN

David S. Brée (bree@cs.man.ac.uk)
Computer Science Department, University of Manchester
Oxford Road, Manchester M13 9PL, U.K.

This poster presents an analysis of the different ways that the preposition IN is used to convey temporal information. IN is the most frequent preposition and so a solution to its semantics will be a good guide to the semantics of other prepositions that are used temporally.

The analysis is based on all 1,980 temporal uses of IN occurring in the Brown corpus (Kučera & Frances, 1967). The temporal use was recognized by the presence of temporal nouns in the IN phrase (Brée & Pratt, 1997). Each occurrence was coded by hand for various features of the phrase, such as determiner, number, qualifier, type of noun, and also for features of the verb in the matrix in which the phrase was embedded such as tense, aspect, modality, negation, superlatives, cardinal and ordinal nouns.

The different possible types of temporal information that IN can convey were found to be:

- a duration (207 occurences) v. an interval (1,653), eg. *in a minute* v. *in 1960*. Almost all (203/207) of the durations were signalled by an indefinite determiner (or no determiner in the case of plural nouns) with a measure type noun, provided the noun was not a qualified plural, nor the noun *time* post-modified. The only other nouns which occurred with an indefinite determiner were life nouns, eg. *youth*, and seasons of the year, but these always indicated an interval.

- a pure duration (58) v. a duration attached to the time of reference (114), eg. *was built in a day* v. *for the first time in 30 years; will report in a day*. Some durations were pure, others were attached to the time of reference. A duration was attached to before the time of reference (39/114) if and only if there was an ordinal, a negative or a superlative in the matrix, generally with a perfect aspect. This is not surprising as the matrix holds for every subinterval in the interval preceding the time of reference but this is not permitted with IN plus a pure duration, FOR being required instead. Generally, a duration was attached to after the time of reference (75/114) if the IN phrase was topicalized (Hitzeman, 1997), if there was a modal in the matrix or if the matrix was a state or achievement rather than an accomplishment.

- an interval attached to a noun phrase v. a verb phrase, eg. *spending in the Sixties* v. *spent in the Sixties*. NP attachment (322) occurred whenever the IN phrase was before the verb, when the matrix was a superlative, when there was some other temporal adverbial, etc.

- measurement (104) v. quantification over an interval (1,226), eg. *sold 7 tractors in August* v. *died in August*. For measure use, the matrix must be either a repeatable event or an activity with repeated outcomes.

- universal (260) v. existential (966) quantification, eg. *accepted in the Fifties* v. *born in the Fifties*. As IN cannot be used for giving the duration of a state or activity, it is surprising for it be used when a matrix state or activity was lasting the whole of an **interval**.

- unique (1,272) v. generic (152) durations and intervals, eg. *made in the 1940s* v. *would come home in the morning*. A generic use was indicated by a quantifying determiner, a season with no determiner, a plural part-of-day noun and sometimes when the tense in the matrix was the simple present, unless this tense was being used to describe a current attitude to a future event, or in the context of art criticism or history, etcetera.

The last four distinctions depend on general features of the matrix and not on the IN phrase at all.

These heuristics provide a means for extracting the temporal semantics of IN phrases that could be implemented in a natural language computer system used, for example, either to understand natural language, or to give deep machine translation into another language. They will generalize to temporal uses of other prepositions, eg. FOR, and also to other uses of IN, in particular spatial uses, which closely parallel the temporal use.

References

Brée, D.S. & Pratt, I.E. (1997). Using prepositions to tell the time. *Proceedings of the Nineteenth Cognitive Science Conference* (pp. 873). Hillsdale, NJ: Erlbaum.

Hitzeman, J. (1997). Semantic partition and the ambiguity of sentences containing temporal adverbials. *Journal of natural language semantics, 5*, 87-100.

Kučera, H. & Frances, W.N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Thoughts on the Prospective MML-TP: A Mental MetaLogic-Based Theorem Prover

Selmer Bringsjord & Yingrui Yang
selmer@rpi.edu • yangyri@rpi.edu
Dept. of Philosophy, Psychology & Cognitive Science (S.B. & Y.Y.)
The Minds & Machines Laboratory (S.B. & Y.Y.)
Department of Computer Science (S.B.)
Rensselaer Polytechnic Institute
Troy NY USA

Why do automated reasoning systems fail to even *hint* at the power of a human mathematician or logician? (See, e.g., the comparison between the "real" Gödel and machine versions of proofs of his famous incompleteness theorems: Bringsjord 1998.) Why is it that the best machine-proved theorems are shallow when matched against minds? From the standpoint of psychology and cognitive science, this is an exceedingly difficult question to answer, in no small part because these fields don't offer a mature theory of mature reasoning. What they *do* offer are theories of *im*mature reasoning, that is, theories of — as it's said — untrained or ordinary reasoning. (We have no precise definition of the difference between expert and novice reasoning, of course; but the distinction seems to make intuitive sense, and we can resort to characterization by example. E.g., *reductio ad absurdum* is probably *generally* the province of expert deductive reasoners. Certainly, say, diagonalization is the province of such reasoners.) If we assume that these theories can scale to the expert case, then it's no surprise at all that automated reasoning systems, when stacked against mathematical minds, look pretty bad. The reason is that theories of ordinary reasoning, before the advent of our Mental MetaLogic, invariably put their eggs in one basket. For example, the representational system of current mental logic theory can be viewed as (at least to some degree) a psychological selection from *only* the syntactic components of systems studied in modern symbolic logic. For example, in mental logic *modus ponens* is often selected as a schema while *modus tollens*[1] isn't; yet both are valid inferences in most standard logical systems. Another example can be found right at the heart of Lance Rips' system of mental logic, PSYCOP, set out in (Rips 1994), for this system includes conditional proof (p. 116), but *not* the rule which sanctions passing from the denial of a conditional to the truth of this conditional's antecedent (pp. 125–126). So, a theorem prover based exclusively on mental logic would of necessity fail to capture human mathematical reasoning that is "semantic." And while it's certainly true that expert reasoners often explicitly work within a system that is purely syntactic, it's also undeniable that such reasoners often work on the semantic side. Roger Penrose has recently provided us with an interesting example of semantic reasoning: He gives the case of the mathematician who is able to see via an image of the sort that is shown in Figure 1 that adding together successive hexagonal numbers, starting with 1, will always yield a cube.

Now, as a matter of fact, nearly all theorem provers are essentially incarnations of mental logic, so it is unsurprising that such provers are impoverished relative to trained human reasoners. Mental MetaLogic (MML) (Yang & Bringsjord under review) is a theory of reasoning which draws from the proof theoretic side of symbolic logic, the semantic (and therefore diagrammatic) side (and hence owes much to Johnson-Laird's 1983 mental model theory), and the content in between: metatheory. Furthermore, while theories of reasoning in psychology and cognitive science have to this point been restricted to elementary reasoning (e.g., the propositional and predicate calculi), MML includes psychological correlates to modal, temporal, deontic, conditional, ... logic. Accordingly, MML-TP will be theorem prover that ranges over a diverse range of representational schemes and types of reasoning.

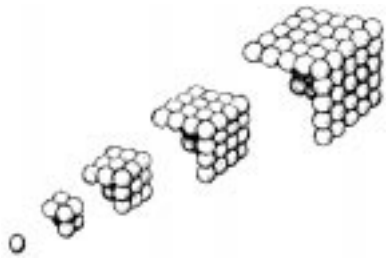


Figure 1: Viewed a Certain Way, the Cubes are Hexagons

[1] From $\phi \rightarrow \psi$ and $\neg\psi$ one can infer to $\neg\phi$.

References

Bringsjord, S. (1998), 'Is Gödelian model-based deductive reasoning computational?', *Philosophica* **61**, 51–76.

Johnson-Laird, P. N. (1983), *Mental Models*, Harvard University Press, Cambridge, MA.

Rips, L. (1994), *The Psychology of Prooof*, MIT Press, Cambridge, MA.

Yang, Y. & Bringsjord, S. (under review), 'Mental metalogic: A new paradigm in psychology of reasoning'.

Hemispheric Effects of Concreteness in Pictures and Words

Daniel J. Casasanto[1] (dcasasan@mail.med.upenn.edu)
John Kounios[2] (jkounios@cattell.psych.upenn.edu)
John A. Detre[1] (detre@mail.med.upenn.edu)
Department of Neurology[1] and Psychology[2], University of Pennsylvania
3400 Spruce Street, Philadelphia, PA 19104 USA

Introduction

Functional Magnetic Resonance Imaging (fMRI) studies have demonstrated differently lateralized activation during episodic memory encoding for verbal and pictorial stimuli, as predicted by the material-specific model (1). Results are interpreted as consistent with Dual Coding Theory, which posits parallel verbal and imaginal systems by which a given stimulus may be represented (2), the neural substrates of which have also been shown to be differently lateralized (3). Whereas previous encoding studies have examined variation across material types, the present study investigated hemispheric effects within material types. FMRI was used to determine the laterality of encoding-related activation for verbal stimuli that varied in concreteness, and for pictorial stimuli that varied in verbalizability.

Methods

Task Design

Verbal task. Nine healthy, right-handed native English speakers viewed blocks of serially presented English nouns, alternating with blocks of control strings (2500ms presentation, 500ms ITI). Noun stimuli comprised two sublists: 40 concrete and 40 abstract nouns (mean concreteness ratings 6.31 and 2.61, respectively, on a 1-to-7 point scale (Toglia & Battig, 1978)). Each task block comprised ten nouns: eight from one sublist and two from the other. Subjects classified nouns as concrete or abstract, responding via a left-or-right button press. Control blocks comprised ten strings of Ls or Rs, eliciting the same proportion of left and right button presses as the preceding block of nouns. Subjects were instructed to remember noun stimuli for a post-scan forced-choice recognition memory test.

Pictorial task. For each of two face memory encoding tasks, eleven healthy, right-handed volunteers viewed blocks of unfamiliar face photographs, alternating with blocks of a repeatedly presented pixelated control image (six 40s task/control blocks, 10 stimuli per block, 3500ms presentation, 500ms ITI). For the first task, full-head photographs were shown, including hair, neck, and upper shoulders. In some cases, clothing and jewelry were visible. For the second task, the same set of face photographs was used, but each photograph was cropped so as to include the brow, eyes, nose, and mouth, but exclude ears, hair, and any extraneous objects. Subjects were instructed to remember the faces for a post-scan recognition test, and to attend the control images but not to memorize them. Scanning occurred during the encoding tasks but not during recognition testing.

Image Acquisition and Data Analysis

BOLD functional images were obtained at 1.5Tesla in 20 contiguous 5-mm-thick axial slices. Multisubject SPMs were constructed in Talairach space using the SPM99{t} random effects model. Cognitive subtraction revealed activation associated with encoding during blocks of predominantly concrete and predominantly abstract nouns, and during cropped-face and full-head encoding. For each task, activation exceeding a statistical threshold ($\propto$=.05) was quantified in two *a priori*-defined regions of interest. ROIs comprised the inferior frontal gyrus (IFG) and fusiform gyri (FG), as these structures have demonstrated reliable material-specific effects in previous encoding studies. Hemispheric asymmetry of activation in each ROI was assessed using an asymmetry ratio [AR = (VoxelsR–VoxelsL)/(VoxelsR+VoxelsL)].

Results and Discussion

Differing hemispheric effects shown previously across verbal and pictorial material types were demonstrated in the present study within material types. Average activation in the ROIs was bilateral during encoding of concrete nouns [AR(IFG)=-.06, *ns*; AR(FG)=.13, *ns*], which are amenable to both verbal an imaginal coding, but significantly left-lateralized during encoding of abstract nouns [AR(IFG)=-.22, P=.001; AR(FG)=-.80, P=.009], which are resistant to imaginal coding. Activation during encoding of full-head photographs was left-lateralized in the IFG [AR(IFG)=-.46, P=.001] and bilateral in the FG [AR(FG)=-.25, *ns*], but activation during encoding of cropped-faces, which are resistant to verbal coding, was significantly right-lateralized in both ROIs [AR(IFG)=.12, P=.001; AR(FG)=.71, P=.001]. Replication of these findings within verbal stimuli that vary in imageability and within pictorial stimuli that vary in verbalizability would suggest that hemispheric specialization during memory encoding heretofore described as material-specific might be more accurately described as code-specific.

References

Casasanto, D.J., *et al.* (2000) in *Proceedings of the Cognitive Science Society* **22**, 77-82.

Paivio, A. (1991) *Canadian Journal of Psychology* **45,** 255-287.

Kounios, J. & Holcomb, P.J. (1994) *J. Exp. Psych.: Learning, Memory, & Cognition* **20,** 804-823.

Learning Statistics: The Use of Conceptual Equations and Overviews to Aid Transfer

Richard Catrambone (rc7@prism.gatech.edu)
Georgia Institute of Technology, School of Psychology,
274 5th Street, Atlanta, GA 30332-0170 USA

Robert K. Atkinson (atkinson@ra.msstate.edu)
Mississippi State University, Department Counselor Education and Educational Psychology,
Box 9727, Mississippi State, MS 39762 USA

Learners can carry out new procedures or solve new problems that are quite similar to those on which they were trained, but they have difficulty when the novel cases involve more than minor changes from what they had previously studied. This transfer difficulty seems to stem from a tendency by many learners to memorize the steps of how equations are filled out rather than learning the deeper, conceptual knowledge that is implicit in the details. One type of knowledge structure that appears to aid procedural generalization is one organized by subgoals (Catrambone, 1998). A subgoal represents a meaningful conceptual piece of an overall solution procedure and can serve as a guide to which part of a previously-learned solution procedure needs to be modified for a novel problem.

In the domain of mathematical problem solving there are often "computationally-friendly" solution approaches in which multiple solution steps are collapsed into a single formula. These formulas allow for the easy calculation of the solution by simply inserting the correct values into the formula. A major drawback of this approach is that it is restricted to solving a narrow range of problems that fall into predefined problem categories corresponding to solution formulas.

A good example of such a contrast is the process of calculating sum of squared deviation scores (SS) for the variance terms in t-tests and analyses of variance (ANOVAs). The conceptual formula for SS in a t-test, $\Sigma\ (X - \overline{X})^2$ translates directly into the sum of squared deviations. This clearly captures how the variance term measures the amount of spread about the mean. In contrast, the computational formula for SS, $\Sigma X^2 - (\Sigma X)^2 / N$ permits the learner to calculate SS directly from raw scores which can lead to more efficient calculations, however, the computational formula conceals the notion of spread.

We were also interested in the relative effectiveness of overview information presented to learners either before they studied an example or after they studied an example. Overview information offered as a pre-instructional aid might, like advance organizers, provide an organizing cognitive structure for receiving new material (Ausubel, 1968). However, a pre-example overview runs the risk of appearing too abstract and non-contextualized. A post-example overview might supplement or reinforce what was illustrated in the examples, thereby facilitating the integration and application of recently acquired procedural knowledge.

Procedure and Hypothesis

Participants (N = 112) studied a t-test example and a two-group ANOVA example that used either a conceptual or computational approach. Overview information either preceded or followed each example. Participants then solved two near transfer problems (a t-test problem and a two-group ANOVA problem) and a far transfer problem (a three-group ANOVA problem that required an adaptation of how variance was calculated). We hypothesized that consistent with a subgoal-learning approach, learners who studied examples with conceptually-oriented equations would transfer more successfully to novel problems compared to learners who studied examples using computationally-oriented equations.

Results and Discussion

The conceptual group outperformed the computational group on the 2-group ANOVA problem; this is a bit surprising because all participants studied a 2-group ANOVA example. One possible explanation for this finding is that the conceptual equations are easier to implement and to check for errors. The conceptual group was also more successful on the 3-group ANOVA problem which is consistent with the hypothesis that these participants were better able to modify the equation.

There was no effect due to the position of the overview. This suggests several possibilities including: the material simply was not very effective; positioning of such material does not matter for learning in this domain; learners may have been less inclined to pay attention to it because the overview material essentially recapitulates the information provided in the examples (with a bit of elaboration) but without numbers.

Brief References and Acknowledgments

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. New York: Holt, Rinehart & Winston.

Catrambone, R. (1998). The subgoal learning model: Creating better examples so that students can solve novel problems. *Journal of Experimental Psychology: General, 127* (4), 355-376.

This research was supported by AFOSR Grant F49620-98-1-0362 to Richard Catrambone.

Infants' Associations of Words and Sounds to Animals and Vehicles

Eliana Colunga (ecolunga@cs.indiana.edu)
Computer Science Department; Lindley Hall 215
Bloomington, IN 47405 USA

Linda B. Smith (smith4@indiana.edu)
Department of Psychology; 1101 East Tenth Street
Bloomington, IN 47405 USA

In a recent study, Woodward and Hoyne (1999) showed that 13-month-olds readily associate both words coming from the experimenter's mouth and non-linguistic sounds coming from a hand-held noisemaker with object categories. In contrast, 20-month-olds associate words but not non-linguistic sounds with object categories. Woodward and Hoyne suggest that words become privileged as possible names; that the forms a name can take are open at the beginning and become more restricted with development. Are children learning what forms count as words? If so, just what defining features are they learning?

In the research presented here, we attempt to answer these questions. In our account, words become privileged as names because of the special way they correlate with categories.

By our account, there are two parts to this specialness. First, there is one name (more or less) that goes with each category (more or less). So, the name of the category is a feature that all members of the category have in common while at the same time the name is a feature that distinguishes instances from members of other categories. Second, words *as a domain* have this special function of pointing to categories. The fact that there are many words that point to categories is what helps children generalize this expectation to novel words. Thus, our account makes two predictions:

1. Any strongly correlated feature of a name will become an integral part of what is a name. Words emanating from mouths is a highly systematic property of names. Thus, we predict that if a word comes out of a place other than a mouth, young children will not take it as a name.

2**.** Any event domain that systematically predicts category membership will be taken as a name as well. For example, animal category correlates with animal sound: dogs bark, cats meow, elephants trumpet and so on. Thus animal sounds should be taken as names for animals.

In Experiment 1 we test this predictions by looking at children's associations of words, animal sounds and motor sounds to animal categories. In Experiment we look at their associations with vehicle categories.

Experiment 1

Thirty-six 20-26 month-old children participated in a 3x3 mixed design with three different sound sources (mouth, animal, noisemaker) as between-subject variable and three different sounds (word, animal sound, motor sound) as within-subject variable. Thus, each child heard three different kinds of names for three different animal categories. In the Mouth condition, children heard the three kinds of label coming from the experimenter's mouth. The experimenter named the animal "Look at this toma" in the Word condition, "Look at this <frog clucking>" in the Animal Sound condition, and "Look at this <motor>" in the Motor Sound condition. In the Animal condition the three kinds of label came from recorders concealed in the toy animal. In the Noisemaker condition the labels emanated from hand-held recorders. The carrier phrase was always said by the experimenter.

During the test phase, children were presented with the target object and a distracter on a tray. The child was asked to "Get the <label>". The results show that 20-26 month-olds 1) associate both words and motor sounds to animal categories only when the words emanate from the mouth and 2) associate animal sounds with animal categories regardless of the source of the sound. In other words, the three kinds of labels coming from the mouth are taken as names, and animal sounds coming from any of the three sources are taken as names.

Experiment 2

The results of Experiment 1 match our predictions. However, an alternative explanation is that there is something inherently word-like about animal sounds. In Experiment 2 we explore this possibility by looking at children's learning of labels for vehicle categories. If babies in Experiment 1 linked the animal sound to the animal toy because of some perceptual property of the sound, we would expect them to link animal sounds to vehicle categories as well. However, if our account is correct, they will reject the animal sound as a label for a vehicle but accept the motor sound.

We used the same design but the stimuli used were novel vehicles. Our results show that, as predicted, 20-26 month-olds take motor sounds as labels for vehicle categories, but reject animal sounds as labels for vehicle categories.

Why this pattern? Because this pattern reflects the systematicity with which events correlate with categories in the world. Sounds from mouths typically name things; animal sounds correlate with animal categories in much the same way as words correlate with object categories. In conclusion, we suggest that it is the systematicity of prior learned pairings that determines what counts as a name.

A Connectionist Model of Semantic Memory: Superordinate structure without hierarchies

George S. Cree (gcree@uwo.ca)
Department of Psychology, 1151 Richmond Street
London, Ontario N6A 5B8 Canada

Ken McRae (kenm@uwo.ca)
Department of Psychology, 1151 Richmond Street
London, Ontario, N6A 5B8 Canada

Symbolic, spreading-activation models of semantic memory represent subset-superset relationships among concepts as distinct, hierarchical levels of nodes connected by "isa" links (e.g., Quillian, 1968). Numerous theoretical and empirical arguments have been leveled against this approach (e.g., Dean & Sloman, 1995; Rumelhart & Todd, 1993), including (1) the difficulty such models have in accounting for familiarity and typicality effects, (2) that category membership is often unclear, (3) that items can belong to multiple categories, (4) that some categories are more internally coherent than others, (5) that general properties do not necessarily take longer to verify than specific properties, and (6) that some general category membership relations can be verified faster than specific category membership relations.

We present a novel connectionist model of semantic memory that offers potential solutions to these problems. The model, an extension of McRae, de Sa & Seidenberg's (1997) and Cree, McRae & McNorgan's (1999) models of semantic memory, was trained to compute distributed patterns of semantic features from word forms. Semantic feature production norms were used to derive basic-level representations and category membership for 181 concepts taken from McRae et al's (1997) property norms. Basic-level (e.g., dog) and superordinate-level (e.g., animal) concepts were represented over the same set of semantic features.

The training scheme was designed to mimic the fact that people sometimes refer to an exemplar with its basic-level label, and sometimes with its superordinate-level label. Two types of training trials were used. In 90% of the training trials, basic-level word forms mapped to their semantic representation, instantiating a one-to-one mapping. The occurrence of each of the 181 basic-level exemplars during training was scaled by familiarity ratings that were collected from human participants. In the remaining 10% of the trials, a superordinate word form was trained by pairing it with one of its exemplars' semantic representations. Importantly, each semantic representation included in a category was paired with that superordinate word form with equal frequency (i.e., typicality was not built in).

The model was used to simulate data from typicality, superordinate-exemplar priming, and category-verification experiments. In explaining the human data, emphasis was placed on the role of correlations among features, the familiarity of concepts, category size, and on the distinction between off-line and on-line processing dynamics. Specifically, settled attractor states for superordinate-level concepts are composed of a greater number of units with states on the linear component of the sigmoidal activation function, making it easier, for example, for the network to move from a superordinate representation to any other during temporal, on-line processing.

Acknowledgments

This work was supported by an NSERC Postgraduate Fellowship to the first author and NSERC grant RGPIN155704 to the second author.

References

Cree, G.S., McRae, K. & McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. Cognitive Science, 23, 371-414.

Dean, W. & Sloman, S.A. (1995). A connectionist model of semantic memory. Unpublished Manuscript.

McRae, K., de Sa, V.R. & Seidenberg, M.S. (1997). On the nature and scope of featural representations of word meaning. Journal of Experimental Psychology: General, 126, 99-130.

Quillian, M.R. (1968). Semantic Memory. In M. Minsky [Ed.], Semantic Information Processing (pp. 216-270). Cambridge, MA: MIT Press.

Rumelhart, D.E. & Todd, P.M. (1993). Learning and connectionist representations. In D.E. Meyer and S. Kornblum [Eds.], Attention and Performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience (pp. 3-30). Cambridge, MA: MIT Press.

Concept Generalization in Separable and Integral Stimulus Spaces

Nicolas Davidenko (ndaviden@psych.stanford.edu)
Joshua B. Tenenbaum (jbt@psych.stanford.edu)
Department of Psychology
Stanford University, Stanford, CA

Many models of human concept learning are built around a hypothesis space of possible concepts with an associated probability distribution (Tenenbaum, 1999). These hypothesis spaces are difficult to describe, even in the case of two-dimensional stimulus spaces. Garner (1974) distinguished two types of stimulus spaces: separable (where similarity judgments follow a city-block metric) and integral (where they follow a Euclidean metric). To explain why generalization contours from one exemplar are diamonds in separable spaces and circular in integral spaces, Shepard (1987) suggested two corresponding hypothesis spaces: the space of axis-aligned rectangles (for separable spaces), and the space of circular discs (for integral spaces). However, the generalization contours from one exemplar do not uniquely determine a hypothesis space (e.g. the hypothesis space of squares under all possible rotations produces the same generalization contours as the hypothesis space of circular discs). In this work, we attempt to constrain the choice of hypothesis spaces by analyzing concept generalization from multiple exemplars. Our preliminary findings are generally consistent with Shepard's formulation except for one significant difference: people show correlational concept generalization even in separable spaces.

Method

Fourteen subjects (ages 15 to 49) participated in a two-part experiment. Each part included 6 trials, and the order of the parts was counterbalanced across subjects. In each trial in part I (concept generalization in a separable space), subjects observed on a computer screen a group of five stimuli said to be "representative of a larger set." The stimuli were circles of variable size with radial lines of variable orientation. The parameters were chosen to fall on line segments embedded in the Size-Orientation space and were either axis-aligned (varying only in one dimension) or correlational (varying in both dimensions simultaneously). The five exemplars were evenly spread in the stimulus space. After observing the exemplars, subjects rated each of eight test items according to the perceived probability that it belonged to the same set represented by the exemplars, on a 0-10 scale. The test items were placed either on the linear extension of the concept or perpendicular to it. Part II (concept generalization in an integral space) followed the same design except that the stimuli were discs varying in two integral color dimensions: saturation and brightness.

Results and Discussion

Figure 1 shows the predictions using Shepard's hypothesis spaces of the 50% generalization contours for axis-aligned and correlational stimulus sets (dark dots).

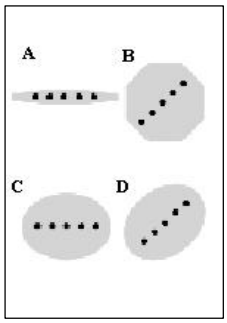


Figure 1

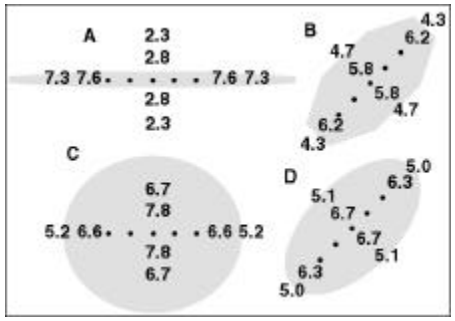


Figure 2

Generalization in a separable space (top row) is orientation-dependent; in particular, adaptation to the one-dimensional extent of the concept occurs only when the concept is axis-aligned (A). In an integral space (C and D), concept orientation is irrelevant.

Figure 2 summarizes the experimental data. Probability ratings of the test items (averaged across subjects) are shown as numbers. The contour regions were obtained by extrapolating the 50% probability boundary. In the separable space (top row), linear adaptation occurs more noticeably in the axis-aligned concept (A), but also occurs to some extent in the correlational concept (B), contrary to an axis-aligned rectangular hypothesis space. This finding suggests that the hypothesis space for separable dimension stimulus spaces is more complex than originally formulated by Shepard, perhaps including rectangular regions of all possible orientations weighted by some prior probability distribution. We are currently exploring this possibility through human experiments and mathematical modeling.

References

Garner, W. R. (1974). The processing of information and structure. Experimental Psychology Series, Potomac, Maryland, Lawrence Erlbaum Associates.

Shepard, R. (1987). Towards a universal law of generalization for psychological science. Science, 237, 1317-1323.

Tenenbaum, J.B. (1999). Bayesian modeling of human concept learning. Advances in Neural Information Processing Systems, 11.

Linguistic Resources and " Ontologies " across Sense Modalities
A Comparison between Color, Odor, and Noise and Sound.

Dubois Danièle (daniele.dubois@inalf.cnrs.fr)
CNRS/LCPE, 44 rue de l'Amiral Mouchez, F-75014 Paris
&
Cance Caroline (caroline.cance@ivry.cnrs.fr)
Université de Paris 3 & LCPE, 44 rue de l'Amiral Mouchez, F-75014 Paris

After exploring categorization of color and other visual objects (Dubois, 1991; 1997), we have been recently investigating cognitive categories within other senses, such as olfaction and audition (Dubois, 1997b). We present here contrasted results concerning psychological representations of colors, odors, noises and sounds induced from their linguistic expression in (French) language.

Quantitative evaluations of occurrences, their morphological, syntactic and semantic properties were computed on a corpus of 108 definitions produced by native speakers, according to previous analyses theoretically based in Dubois (2000; Dubois & Grinevald, 1999). Only partial results and conclusions will be reported here.

Table 1: Number of nominal forms: Number of occurrences and number of single occurrences (Hapax) in the 4 corpora:

	forms	Occ.	Hapax
odor	78	158	54
color	67	117	48
sound	20	32	14
noise	14	24	10

Among other indicators such as the number of verbs, relative clauses, adjectival forms (simple or deverbal ('pleasant') or denominal ('noisy') constructions), the nominal forms reveal that French linguistic resources vary across sense modalities : acoustic representations show less productivity and more agreement between subjects than colors and than odors.

Table 2: Linguistic marks of "objectivity" and personal involvement (Percentage of subjects producing the word "*something*" and personal pronouns in their definitions)

	"something"	personal pronouns
odor	24,3	43,9
sound	25,9	35,2
noise	20,4	33,3
color	14,8	21,4

The lack of commonly shared naming for odors and acoustic phenomena correlates with the uncertainty of their definitions, stated as "something" that affects the subject, as reflected in the greater personal involvement for odor than for sound and noise, and lesser for color definitions.

Conclusion

Colors as visual objects seem to be processed as stimuli "standing out there", whereas odors are more likely to be structured as **effects** of the world on the subject, therefore less autonomous from the experiential context. Acoustics phenomena can be represented at different degrees of "subjectivity" (or objectivity), contrasting noises that are more subjective than sounds, these latter referring to a more expert, objective, technical and scientific knowledge.

If we always perceive "something", through the diversity of senses, language diversely objectivizes and "stabilizes" our cognitive representations of the world into a large variety of linguistic forms. These forms may constrain the "ontology" given to the entities and lead to different distances between the "subject" and the "objects" of the world, from complex phrasing expressing the effects of the world on the subject, to simple " basic " names, which suggest the idea that things are "crying out to be named".

References

Cance, C. (2000) *Définitions d'objets sensoriels en langue française : odeurs, couleurs, bruits et sons.* DEA de sciences cognitives, Université de Lyon 2.

Dubois D. (Ed) (1991). *Sémantique et Cognition*. Paris: Ed. du CNRS.

Dubois, D. (1997). Cultural beliefs as non-trivial constraints on categorization: evidence from colors and odors. *B.B.S*,. 20, 2, 188.

Dubois, D. (Ed.) (1997)b. *Catégories et cognition, de la perception au discours* Paris : Kimé.

Dubois, D., & Grinevald, C. (1999) "Denominations of colors in practices", *XXVI LACUS forum proceedings,* (Edmonton, 1999), pp. 237-246.

Dubois, D. (2000) Categories as acts of meaning, *Cognitive Science Quartely*, 1, 35-68.

What was the Cause? Children's Ability to Categorize Inferences

(ellefson@siu.edu)
Department of Psychology, Mailcode 6502
Southern Illinois University - Carbondale
Carbondale, IL 62901-6502 USA

Inferences have been utilized in a number of studies to further investigate the dynamics of comprehension in children (e. g., Casteel, 1993). An inference can be defined as the processing of information that extends beyond the initial processing of text (Inman & Dickerson, 1995). An inference is usually made when information is not specifically indicated in the text (McKoon & Ratcliff, 1992). An inference is a critical part of comprehension (e.g., Oakhill, 1984; Phillips, 1988). Often causal relations are not specified and an inference is needed. Children draw inferences with relative ease and the study of inferences has helped researchers understand the dynamics of comprehension in children. In previous research children have been able to generate their own responses from ambiguous text (e.g., Bonitatibus & Beal, 1996; Casteel, 1993). The current study examines children's ability to classify inferential statements.

Nineteen fourth-grade participants (9 males and 10 females) between the ages of 9 and 10, read stories with ambiguous endings. Each story was followed by 6 statements that elicited either an unlikely, neutral, or likely inference. After each statement, the children indicated, using a "Yes" or "No" key, whether they thought the statement could fit into the context of the story.

Responses and reaction time data were collected from each participant. There was a main effect of sentence type (F (2, 36) = 5.43, $p < .05$). Reaction times were significantly faster to the unlikely statements compared to both the likely (F (1,38) = 10.28, $p < .01$) and neutral statements (F (1, 36) = 5.13, $p < .05$).

Overall, children correctly classified unlikely inferences as not appropriate for the story (87%). Most children also rated the neutral inferences as not fitting into the story. Children were reluctant to classify likely inferences as fitting into the story. Overall, only 2/3 of the likely statements were classified as fitting into the story. The lower classification rate for the likely sentences was significant (F (2, 87) = 173.32, $p < .001$). The children's reluctance to classify likely sentences as fitting into the story is reflected by significant difference in reaction time between the "Yes" and the "No" responses (F (1, 88) = 16.73, $p < .001$).

The results indicated that children rejected unlikely inferences much faster than they accepted likely or neutral inferences. The children may have generated their own inferences during the story. After the story they had to compare their prior expectations with the likely, neutral, and unlikely inferences. If children have generated correct inferences than unexpected statements like those in the unlikely condition may have been quickly rejected. The additional time used to classify the neutral statements may have been necessary because the statement could have happened, but often did not fit with the context of the story. The likely statements may have been close to the children's expectations and would require additional processing to be sure that they really did fit with the story. Children were reluctant to respond "Yes", which might have been a strategy to ensure their accuracy. This research indicates that further work is necessary for a better understanding of how children process ambiguous information.

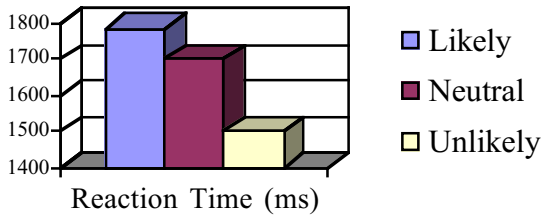


Figure 1. Reaction time (ms) for classification of Likely, Neutral, and Unlikely inferences.

References

Bonitatibus, G. J., & Beal, C. R. (1996). Finding new meanings: Children's recognition of interpretive ambiguity in text. *J. of Exper. Child Psych.*, *62*, 131-150.

Casteel, M. A. (1993). Effects of inference necessity and reading goal on children's inferential generation. *Devel. Psych.*, *29*, 346-357.

Inman, T., & Dickerson, D. (1995). Children's inferences of causal events during story understanding. *J. of Genetic Psych.*, *156*, 265-277.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psych Rev, 99*, 440-466.

Oakhill, J. (1984). Inferential and memory skills in children's comprehension of stories. *Brit J of Educ Psych, 54*, 31-39.

Phillips, L. M. (1988). Young readers' inference strategies in reading comprehension. *Cognition and Instruction*, *5*, 193-222.

Structural Alignment in Similarity and Difference of Simple Visual Stimuli

Zachary Estes (zcestes@princeton.edu)
Uri Hasson (hasson@princeton.edu)
Department of Psychology, Princeton University
Princeton, NJ 08544-1010, USA

The present investigation tested the predictions of Structural Alignment theory (Gentner & Markman, 1997) in similarity and difference judgments of simple visual stimuli. Alignment theory explains comparison as a process of aligning the structure of one stimulus with the structure of the other stimulus. The theory makes a critical distinction between alignable differences, which are related to commonalities in the structures of two stimuli, and nonalignable differences, which have no structural correspondence in the two stimuli. Examples are shown in Figure 1. The difference between the standard stimulus S and the NAD stimulus is nonalignable because S has no element that corresponds to the black triangle in NAD. The difference between S and the ADA stimulus, on the other hand, is alignable because the white circle in S is aligned with, or corresponds to, the black circle in ADA. Alignable differences may also occur in the form of a different object (ADO) or a different relation between objects (ADR).

Figure 1: Stimuli used in Experiments 1 and 2.

These distinctions are critical for predicting similarity and difference judgments. Alignment theory predicts that "alignable differences count more against similarity than nonalignable differences" (ibid, p. 50). That is, items with an alignable difference (i.e., ADA, ADR, and ADO) should be judged less similar to (and more different from) the standard than should items with a nonalignable difference (i.e., NAD). A second prediction is that the more different the alignable difference is from the standard, the more it will detract from similarity (see Markman & Gentner, 1996).

Experiments 1 and 2. Stimuli consisted of all possible pairs of items shown in Figure 1 (excluding the standard stimulus), thus creating 6 item pairs. For each item pair, participants judged which of the two stimuli was more similar to the standard stimulus (Experiment 1) or more different from the standard stimulus (Experiment 2).

Table 1: Proportions of similarity and difference choices.

Item pair	Similarity	Difference
(1) ADA & ADO	ADA = .65	ADO = .70
(2) ADR & ADO	ADR = .55	ADO = .58
(3) ADA & ADR	ADA = .74	ADR = .57
(4) NAD & ADA	ADA = .74	NAD = .72
(5) NAD & ADR	ADR = .69	NAD = .68
(6) NAD & ADO	ADO = .53	NAD = .50

Discussion. Comparisons (1) and (2) in the Table above show that, of the items with alignable differences, ADO was most different from S. Comparison (3) shows that ADA was the least different from S, with ADR falling in between. Having established this hierarchy of alignable differences, we next examined whether the degree of difference of an alignable difference from S did affect the degree to which that alignable difference detracted from similarity (when judged with a nonalignable difference). As predicted, comparison (4) shows that the least different alignable difference detracted the least from similarity judgments (i.e., ADA = .74), while (6) shows that the most different alignable difference detracted the most from similarity judgments (i.e, ADO = .53). These findings extend and replicate those of Markman and Gentner (1996).

However, as apparent in the Table, in no case did an alignable difference (i.e., ADA, ADR, or ADO) detract more from similarity judgments than did a nonalignable difference (i.e., NAD). On the contrary, ADA and ADR actually counted less against similarity (and conversely more against difference) than did NAD. This result does not support the prediction of alignment theory.

Potential explanations of this failure to support alignment theory are that (i) alignment theory is not applicable to simple visual stimuli, (ii) the alignable differences used in these experiments were not sufficiently different, or (iii) NAD was not really a nonalignable difference, but rather was an alignable difference in the number of elements in the item. We would be delighted to discuss these and other possibilities with you at our poster.

References

Gentner, D. & Markman, A.B. (1997). Structure mapping in analogy and similarity. American Psychologist, 52, 45-56.

Markman, A.B. & Gentner, D. (1996). Commonalities and differences in similarity comparisons. Memory & Cognition, 24, 235-249.

Music Evolution: The Memory Modulation Theory

Steven D. Flinn (s.flinn@manyworlds.com)
ManyWorlds, Inc., 510 Bering Drive, Suite 470
Houston, TX 77057 USA

Solving the Mystery of Music

The evolutionary basis for the development of advanced musical capabilities in humans has remained a mystery. Whereas language clearly confers obvious fitness advantages, music has resisted such an easy explanation. Current explanations tend to fall into the categories of either music as a by-product of the evolution of other facilities that do directly confer fitness benefits, or music as a sexual selection fitness indicator (Wallin, Merker & Brown, 2000).

We have found neither of these classes of explanations, nor any other of the previously proposed explanations compelling. Rather, we argue that advanced musical capabilities evolved because they directly confer specific fitness advantages. In particular, we argue that human musical capabilities are precisely those set of specialized mental capabilities that co-evolved with language to enable the sophisticated memory modulation of the receiver of information by the communicator of a message. In other words, music enables individuals and groups that are communicating messages to have a degree of control over how the messages will be retained in the memory of the receiver(s). Particularly in the pre-literate world, such abilities have obvious, direct evolutionary advantages.

Memory Modulation and Music

The consolidation theory of memory (McGaugh, 2000) continues to guide current memory research. The theory suggests that it takes time for long-term memories to consolidate. "Considerable evidence suggests that the slow consolidation of memories serves an adaptive function by enabling endogenous processes activated by an experience to modulate memory strength". In other words, it is optimal for long-term memory processes to be highly selective.

Key facilitators of memory modulation are emotional arousal, repetition, and structure. And these are exactly the essential attributes of music -- making it the ultimate vehicle for fine-tuned memory modulation.

In particular, music enhances the probability of long-term memory of coincident events and communications. It appears that music facilitates long-term memory primarily through the evocation of emotion, and with contributions from repetition and additional structure (Schulkind, Hennis & Brown, 1999)

An Information Theoretic Approach

From an information theory point of view, the generalized issue at hand is precisely how communications modes would co-evolve with increasing intelligence, given specific memory storage architectures.

If the architecture of the human brain had been such that there existed only one type of memory, then a communications capability relying simply on syntactical structures would have been sufficient. However, the durability of memories in the human brain (and other animals) can, as a first approximation, be divided into two categories: short-term and long-term memory. Syntactical structures of language alone offer limited ability for the sender of a message to effectively influence the strength of memory of the receiver.

However, the ability of communicators to directly influence strength of memory in receivers would be of exceedingly high fitness value as intelligence and the sophistication of associated messages increased. Indeed, the encoding of a message in very long-term memory significantly increases the probability that the message will be re-transmitted with high fidelity by the original receiver to others. This cascading of the original message vastly increases the evolutionary value of preferential memory selection by the message sender.

We argue, music is, therefore, just that (expected) mode of communication that co-evolved with language and overall intelligence that enabled finer and finer control of the memory modulation of message receivers by message sender(s). Music has all the right characteristics to fit this critical (and expected) role, and there is no other such communications mode that fills such a role as effectively. Nor do we find any other explanation for the evolution of musicality in humans that is as comprehensive and compelling.

References

McGaugh, J. L. (2000). Memory -- a Century of Consolidation. *Science*, *287*, 248-251.

Schulkind, M., Hennis, L., & Rubin, D., (1999). Music, Emotion and Autobiographical Memory. *Memory and Cognition*, *27*, 948-955.

Wallin, N., Merker, B., & Brown, S., (Eds.) (2000). *The Origins of Music*. Cambridge, MA: MIT.

Language affects memory, but does it affect perception?

Michael C. Frank (mcfrank@stanford.edu)
PO Box 15007, Stanford University
Stanford, CA 94309 USA

Lera Boroditsky (lera@psych.stanford.edu)
Department of Psychology, Jordan Hall, Building 420
Stanford, CA 94305-2130 USA

Does the language you speak affect the way you perceive the world? The strong Linguistic Determinism view — the idea that all aspects of thought (even low-level perceptual abilities) are determined by language — is most closely associated with the writings of Benjamin Lee Whorf (1956). Whorf s ideas have generated much interest and controversy, with much of the empirical work focusing on color perception.

Different languages divide the color spectrum differently; does this lead speakers of different languages to actually perceive colors differently? Early studies claimed no differences in color perception (Heider, 1972), but recent cross-linguistic studies (Davidoff et al., 1999) as well as studies of categorical learning (Goldstone, 1994) have claimed that linguistically learned categories can indeed affect people s perception of shapes and colors.

Although it would be exciting to discover effects of language in domains as low-level as the perception of color, we should be careful in establishing what counts as a test of perception. The studies mentioned above may best be characterized as tests of memory, rather than tests of perception. In these studies, subjects were shown a color sample, and then after a delay asked to either select the same color from a pair of alternatives, or to indicate whether a new color sample is the same or different from the one presented previously. Since these tasks rely heavily on subjects ability to remember the color over a delay, they may tell us more about the ability of language to interfere with color memory than color perception. Further, since language may interfere with color memory in trivial ways (e.g., it could act as a secondary code in memory, and thus affect memory performance without altering the actual perceptual memory trace) (Lucy & Shweder, 1979), more research is necessary to establish whether linguistic information is indeed capable of affecting color perception or even perceptual memories.

We have set out to distinguish the effects of linguistically learned distinctions on memory from effects on perception. In one study, we created a continuum of blue color samples, and taught subjects a categorical boundary in the middle of this space of blues (a boundary corresponding to the goluboy/siniy distinction made in Russian). Another group of subjects received comparable exposure to the color samples, but were not taught the categorical boundary.

In the test phase, participants saw pairs of color samples on a screen at the same time and were asked to determine whether the two colors were exactly the same or different as quickly as possible. We were interested in whether learning a category boundary would make people faster to say that two colors are different if they fell in different categories, or slower to say they are different if the two colors were from the same category. Unlike the earlier studies, this task does not rely on subjects color memory since both samples necessary for comparison were presented at the same time. We found no difference in performance between subjects taught the goluboy/siniy distinction and the control group. These preliminary results suggest that color perception may indeed be immune to modification by language. That is, while language may affect color memory, its ability to affect color perception is somewhat doubtful.

Acknowledgments

This research is funded by a URO grant from Stanford University made possible by the generous contribution of Mel and Joan Lane.

References

Davidoff, J., Davies, I., & Roberson, D. (1999). Colour categories in a stone-age tribe. *Nature*, 398, 203-204.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *JEP: General*, 123(2), 178-200.

Heider, E. R. (1972). Universals in color naming and memory. *JEP*, 93, 1, 10-20.

Lucy, J. A., & Shweder, R. A. (1979). Whorf and his critics: linguistic and nonlinguistic influences on color memory. *American Anthropologist*, 81, 581-615.

Whorf, B. L. (1956). *Language, thought, and reality: selected writings of . . .* Carroll, J. B. (Ed.). Cambridge, MA: Technology Press of MIT.

Pragmatic Knowledge and Bridging Inferences

Raymond W. Gibbs (gibbs@cats.ucsc.edu)
Department of Psychology
University of California, Santa Cruz
Santa Cruz, CA95064

Tomoko Matsui (matsui@icu.ac.jp)
Division of Languages
International Christian University
3-10-2 Osawa, Mitaka
Tokyo, 181-8585 JAPAN

What kinds of pragmatic information are necessary for drawing contextually appropriate bridging inferences in conversation (e.g. linking statements such as "I prefer Italy to England. The pasta there was better")? We examined two contemporary pragmatic models of inference generation. One model, Levinson's (2000) view of "presumptive meanings", assumes that inference generation is governed by two default rules that access rich pragmatic knowledge later on in the comprehension process. Another view, Sperber and Wilson's (1995; Matsui 2000), claims that bridging inferences are efficiently generated as implicated premises given the expectation of particular relevant cognitive effects. These models differ, then, in their predictions of when enriched pragmatics shapes utterance interpretation. We report the findings from a series of experiments, measuring both participants' intuitions and reading-times for different bridging inferences. These data generally support the relevance-theoretic view.

The expectation of particular cognitive effects to be achieved by an incoming utterance may not always be generated in the hearer's mind. However, according to Sperber & Wilson, a question is an explicit way of communicating the cognitive effects to be achieved by an incoming utterance. This, in turn, suggests that the person who asks a question is entitled to have rather strong expectations of particular cognitive effects: namely, the relevant answer to his question. One of our experiments, therefore, is designed to compare the comprehension time of two utterance types: a question-answer pair and a juxtaposed utterance pair, each of which describes a state of affairs. Consider the following two sets of utterances, the first of which involve a classic example of bridging inference, namely, the beer was part of the picnic, and the second, a less likely variation:

(1a) Mary: How was the picnic?
John: Well, the beer was warm.
(1b) John: I unpacked the picnic. The beer was warm.

(2a) Mary: How was the job interview?
John: Well, the beer was warm.
(2b) John: I had a job interview. The beer was warm.

A relevance-theoretic view of questions predicts that it is faster to comprehend the utterance 'the beer was warm' in (1a) than (1b). The expected difference in processing time may be explained in terms of how highly accessible the implicature of each utterance is, which possibly facilitates the overall interpretation process. John's utterance in (1), combined with other assumptions and Mary's expectation that John is providing an answer to her question, straightforwardly yields an implicature that the picnic was not totally successful. By contrast, the second utterance in (1b) does not seem to yield any strong implicature, hence, it is predicted that there is no facilitation of the same type. We also predict that the difference in comprehension time is greater between (2a) and (2b) than between (1a) and (1b). In (2), where the relationship between 'job interview' and 'beer' is rather distant, without an expectation of particular cognitive effect, it may almost be impossible to find an acceptable interpretation for (2b). By contrast, in (2a), the interpretation of the second utterance is constrained by the question to the extent that the hearer has to construct the assumption, say, that the beer was offered during the job interview.

Our findings in support of relevance theory suggest that pragmatic information of roughly the same sort enters into listener's understanding of both what speakers say and what they implicate. Bridging inferences appear to be drawn as part of the implicated premises that arise from listeners' attempts to derive appropriate cognitive effects as part of their assumptions about relevance in ordinary communications. These results illustrate the critical importance of enriched pragmatic knowledge in early aspects of linguistic processing.

References

Levinson, S. (2000). *Presumptive Meanings.* Cambridge, Mass: The MIT Press.

Matsui, T. (2000). *Bridging and Relevance*. Amsterdam: John Benjamins.

Sperber, D. & Wilson, D. (1986/95). *Relevance: Comprehension and Cognition*. Oxford: Blackwell.

The AMBR Model Comparison Project: Multi-tasking, the Icarus Federation, and Concept Learning

Kevin A. Gluck (kevin.gluck@williams.af.mil)
Air Force Research Laboratory, 6030 S. Kent St.
Mesa, AZ 85212-6061, USA

Michael J. Young (michael.young@wpafb.af.mil)
Air Force Research Laboratory, 2698 G St., Bldg. 190
Wright-Patterson AFB, OH 45433-7604, USA

Introduction

In recent years, the Human Effectiveness directorate of the Air Force Research Laboratory (AFRL) has increased its investment in science and technology for human behavior representation. One beneficiary of this increase has been the Agent-based Modeling and Behavior Representation (AMBR) Model Comparison Project. The primary goal of the AMBR Model Comparison Project is to advance the state of the art in cognitive and behavioral modeling. It is organized as a series of model comparisons, orchestrated by a moderator team at BBN Technologies. In each comparison, a challenging behavioral phenomenon is chosen for study. Data are collected from humans performing the task. Cognitive models representing different modeling architectures are created, run on the task, and then compared to the collected data. This poster presentation will include results from the first two rounds of the comparison, ongoing work in Round 3, and future plans for Round 4.

Round 1: Multi-tasking

The first iteration of the AMBR Project is complete. The modeling goal in the first round was multi-tasking, and the task domain required a simplified version of en-route air traffic control. Modelers using ACT-R, COGNET/iGEN, D-COG, and EPIC-Soar participated in Round 1. All were able to approximate the trends and central tendencies of the data, but naturally the particular implementation of multi-tasking capability differed across architectures. Round 1 provided a motivation for extending and/or testing each of these architectures in a new way. It was particularly noteworthy that all four utilized some form of "embodiment" (e.g., eyes, hands), although at different levels of fidelity.

Round 2: The Icarus Federation

In Round 2 of the AMBR Model Comparison Project, the Defense Modeling and Simulation Office (DMSO) sponsored the conversion of the simulation environment and models from Round 1, so that they are compliant with DMSO's High-Level Architecture (HLA). Goals for Round 2 include the following:

- Develop an HLA-compliant testbed for research in human behavior representation (HBR)
- Assess the adequacy of the HLA for supporting HBR research
- Assess the adequacy of DMSO's FEderation Development and Execution Process (FEDEP) as a framework for creating and running federations for HBR research

Round 3: Concept Learning

To increase the cognitive requirements of the task used in Rounds 1 and 2, the air traffic control simulation is being supplemented with an embedded category learning task. Multiple aircraft will query the controller (the one that is being modeled) about the possibility of changing altitude. The controller will make a decision to authorize an altitude change based on a multi-dimensional attribute matrix that might include dimensions like aircraft size, level of atmospheric turbulence, and current altitude. The Controller must learn the appropriate responses on the basis of feedback received through the user interface concerning whether they made a correct decision or not. This concept learning task is based on the original laboratory study by Shepard, Hovland, and Jenkins (1961), and modeling studies reported by Nosofsky, et al. (1994).

Round 4: Under Development

The task for Round 4 will be fundamentally similar to the task used in Round 3, but the details are still under consideration. Based on the results of the Round 3 model evaluations, the Round 4 task will be designed to further stress the models and examine their capabilities. We anticipate a focus on the ability of models to adapt from one set of learned concepts to a new, changed set of concepts based on the same or a similar set of concept attributes. Other manipulations such as the workload of the perceptual motor task may also be explored as deemed appropriate given the results of Round III.

References

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: a replication of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, *22*, 352-369.

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 75*(13, Whole No. 517).

Does Adult Category Verification Reflect Child-like Concepts?

Robert F. Goldberg (rgold@pitt.edu)
Department of Psychology and Learning Research and Development Center,
University of Pittsburgh and the
Center for the Neural Basis of Cognition
3939 O'Hara St., Pittsburgh, PA 15260 USA

Introduction

Category verification tasks require subjects to respond as quickly and as accurately as possible as to whether a particular instance is a member of the given category. The variations in reaction time are then used as evidence for the structure of that category. This paradigm led to the view that category structure is based upon graded relations (e.g. Rosch & Mervis, 1975). More recently it has been argued that a theory-based approach is necessary for categorization (Murphy & Medin, 1985). In particular, children are shown to use naïve theories to guide their conceptual development (Carey 1985; Keil, 1989). But surprisingly developmental trends are not usually followed to the structure of the adult endstate (Coley, 2000).

The present research examines the adult endstate of the category living thing. That is, young children tend to conflate animacy, or generalized movement, with alive, while plants are seen as inanimate things and not as biological entities. Reaction times allow a fine-grained analysis to determine if this developmental foundation is present within the adult structure of living thing. The question is whether the acquisition of a well-developed biological theory completely overwrites the misconceptions of childhood.

Method

In each of three experiments, approximately 30 subjects (undergraduates at the University of Pittsburgh) were asked to confirm or deny in a category verification task whether the presented words represent living things. The words were drawn randomly from lists that reflected theoretically motivated contrasts (see below). Relevant lists were yolk matched according to average word frequency, letter length, and number of syllables. Subjects responded to each word three times across 480 trials.

Experiment 1 tested whether adults would be slower and less accurate in denying that animate things (e.g. cloud, car, etc.) are living than they are denying inanimate things (bed, coat, etc.). Experiment 2 predicted that adults would be slower and less accurate in affirming that plants (tulip, elm, etc.) are living things when compared to animals (robin, tiger, etc.). Experiment 3 differentiated the 'no' responses into four word lists: natural animate (ocean, blizzard, etc.), animate (yacht, airplane, etc.), natural inanimate (mountain, pebble, etc.), and inanimate (napkin, desk, etc.) things. It was predicted that instances more similar (e.g. natural animate things) to the category living thing would cause more difficulty in denying membership. Experiment 3 was also expected to replicate the findings of Experiments 1 and 2.

Results and Discussion

The predictions for Experiment 1 were supported. Adults were slower, by about 20ms (Paired samples t-test, $t=3.82$, $p<.001$), and less accurate, by more than 4% ($t=6.14$, $p<.001$), in denying animate things in contrast to inanimate things for the category living thing. In Experiment 2, subjects were slower, by more than 50ms ($t=7.52$, $p<.001$), and less accurate, by almost 10% ($t=8.31$,$p<.001$), in affirming plants than animals. These effects were replicated in Experiment 3. In addition, adults had significantly more difficulty with instances that are highly similar to the category of living things, yet cannot be considered members. Subjects were about 50ms ($t=8.07$, $p<.001$) slower and over 12% less accurate ($t=11.95$, $p<.001$) for natural animate instances than for inanimate things. Adults also had some difficulty denying membership to natural inanimate and animate things.

The results provide an intriguing view into the adult endstate for biological knowledge. It seems that feature associations learned early in childhood remain embedded in the adult structure of the category living thing, even after the formation of a well-developed biological theory. Developmental vestiges would seem to be at the heart of the adult category structure. Future work will be aimed at further clarifying these results in light of traditional views on conceptual change.

References

Carey, S. (1985). Conceptual change in childhood. Cambridge: MIT Press.

Coley, J.D. (2000). On the importance of comparative research: The case of folkbiology. Child Development, 71, 82-90.

Keil, F.C. (1989). Concepts, Kinds, and Cognitive Development. Cambridge: MIT Press.

Murphy, G.L. & Medin, D.L. (1985). The role of theories in conceptual coherence. Psychological Review, 88, 289-316.

Rosch, E. & Mervis, C.B. (1975). Family resemblances: Studies in the internal structure of categories. Cognitive Psychology, 7, 573-605.

Imagining the Impossible

James A. Hampton (hampton@city.ac.uk) and Alan Green
Department of Psychology, City University,
Northampton Square, London EC1V 0HB, UK

Zachary Estes (zcestes@princeton.edu)
Department of Psychology, Princeton University
Princeton, NJ 08544-1010, USA

Research on conceptual combination has followed two main paths. One has examined the representation of logical conjunctions of concepts, such as A FISH THAT IS ALSO A PET. The Attribute Inheritance model posits that conjoint concepts will inherit attributes that are necessary for (or salient in) either of the parent concepts, and will not inherit attributes that are impossible for either of the parent concepts (Hampton, 1987, 1997). The other path has considered the process of combining nonconjunctive concepts, such as ZEBRA CLAM. The Alignment model claims that conceptual combination entails an alignment and comparison of conceptual structures (Wisniewski, 1996). The aim of the current studies was to bridge these two paths by testing whether the alignment and comparison processes ordinarily used in nonconjunctive combination might be the mechanism by which attribute inheritance occurs in conjunctive combination as well.

With most conjunctive combinations, the necessary and impossible attributes of one constituent concept tend to be compatible with the necessary and impossible attributes of the other concept, thus producing a comprehensible combination. Because of this compatibility, unfortunately, such conjunctions do not provide clear evidence either for or against the use of alignment in attribute inheritance. That is, it is unclear whether the attributes are simply inherited by the conjunction, or whether the concepts must first undergo alignment and comparison processes.

One way that alignment and comparison can be observed in attribute inheritance is to present concepts that are incompatible in important respects. Where an attribute is necessary for one concept in a conjunction but impossible for the other, if alignment and comparison occur, then the incompatibility should be detected and somehow resolved. If alignment does not occur, then the incompatibility need not be detected. Thus, we asked participants to imagine the impossible – that is, to conjunctively combine concepts that are in reality disjunctive (e.g., A COMPUTER WHICH IS ALSO A TEACUP).

Study 1

Students were asked to imagine 9 conjunctions (e.g., A FRUIT WHICH IS ALSO FURNITURE) and to describe them in words or pictures. Analysis of the solutions suggested two main findings. First, concepts tended to be instantiated at the basic level (e.g. BANANA for FRUIT, COUCH for FURNITURE). And more importantly, there was strong evidence for alignment and comparison in conjunctive combination. Specifically, the concepts were aligned (e.g., the skin of the banana was aligned with the covering of the couch), conflicting attributes were identified (e.g., bananas rot, couches should not), and emergent attributes were constructed in order to resolve those conflicts (e.g., genetically modified bananas that do not rot). Thus, alignment did indeed appear to be the process by which attributes were inherited in concept conjunction.

Study 2

Given that superordinate classes impose fewer constraints on interpretation than do basic level concepts, one might expect greater success at conjoining superordinate concepts than basic level concepts. On the other hand, superordinate concepts tend not to be alignable with one another. If alignment is necessary for concept conjunction, then superordinates should instead be difficult to conjoin (Markman and Wisniewski, 1997). Study 2 independently manipulated whether the modifier and head concepts in a conjunctive combination were basic (e.g. BANANA) or superordinate (e.g. FRUIT). Solutions were rated by independent judges for their success in terms of conjunctive interpretation. Rated success of solutions did not differ between conditions, suggesting that any advantage superordinates may have had by way of less constraints was offset by their disadvantage of being less alignable as well. Across both experiments, then, there was evidence that attribute inheritance in concept conjunction occurs via alignment and comparison.

References

Hampton, J.A. (1987). Inheritance of attributes in natural concept conjunctions. Memory and Cognition, 15, 55-71.

Hampton, J.A. (1997). Emergent attributes in conceptual combinations. In T.B. Ward, S.M. Smith & J. Vaid, (Eds.), Creative Thought: An Investigation of Conceptual Structures and Processes. Washington DC: American Psychological Association Press.

Markman, A.B. & Wisniewski, E.J. (1997). Similar and different: The differentiation of basic level categories. Journal of Experimental Psychology: Learning, Memory, and Cognition, 23, 54-70.

Wisniewski, E.J. (1996). Construal and Similarity in Conceptual Combination. Journal of Memory and Language, 35, 434-453.

Understanding Negation – The Case of Negated Metaphors

Uri Hasson (uhasson@princeton.edu)
Sam Glucksberg (samg@princeton.edu)
Department of Psychology, Princeton University
Princeton, NJ 08544-1010 USA

Understanding negated statements is more difficult than understanding affirmative ones. People are slower to verify that negative statements are true and seem to understand them best in contexts that make them particularly plausible (see Horn, 1989). Little, however, is known of the online processes that take place during the comprehension of such statements. Particularly, the nature of the representation of negated statements is undecided. Two hypotheses have been raised regarding this issue: according to the first, representation of negation is representation of that which is not the case. This is well summarized by Fauconnier (1994): "negatives set up corresponding counterfactual spaces in which the positive version of the sentence is satisfied". According to the second hypothesis, comprehension of negation goes beyond the comprehension of the affirmative since it involves active inference-making (Manktelow and Over, 1990).

In both cases, negated statements should take longer to process than affirmative ones, since additional cognitive work is required. To explore this issue we presented participants with negated metaphors, e.g., "this train is not a rocket". Following each metaphor, we presented participants with a lexical decision task. The words presented for lexical decision were related either to the negated form of the metaphor (e.g., slow), or to the affirmative form (e.g., fast).

If negation involves processing beyond the affirmative, we would expect that the affirmative meanings of sentences be activated early on, with negation activated only later. Accordingly, we would expect affirmative-related words to be facilitated early on, with negative-related words facilitated only later.

Method

Eighty undergraduate students from Princeton University participated in the experiment for course credit. The variables manipulated in the study were: (a) the time between the endpoint of reading the sentence and the lexical decision task (150, 500 and 1000 ms), (b) the type of prime sentence (Negated metaphor, Affirmative metaphor and Control metaphor) and (c) the type of word presented for lexical decision (related to the negated metaphor or the affirmative metaphor).

Results and Discussion

Figure 1 presents difference from baseline response-times for affirmative- and negative-related target words after reading a negated metaphor. For present purposes, we will not consider response patterns for affirmative sentence primes.

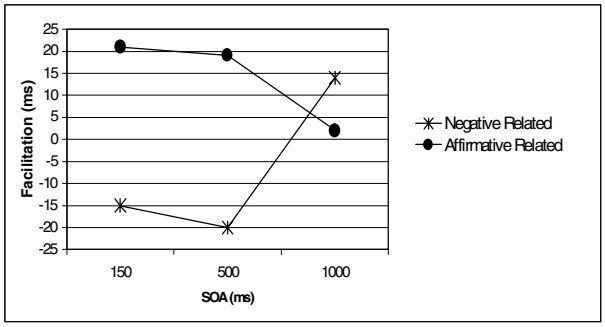


Figure 1: Target facilitation after negated metaphors.

Initially after reading a negated metaphor (e.g., That kindergarten is not a zoo), words associated with the affirmation (i.e., noisy) were facilitated, while words associated with the negation (i.e., calm) were not. This result was also found at 500 msec. By 1000 msec, negative related words were facilitated while affirmative-related words were at baseline level.

This pattern of priming suggests that negated statements are initially represented in their affirmative form. In a short time, however, the affirmative meaning is no longer active, and is replaced by the negated meaning. The findings are consistent with both Fauconnier's, and Manktelow and Over's hypotheses.

References

Horn, L. R. (1989). A Natural History of Negation. Chicago: University of Chicago Press.

Fauconnier, G. (1994). Mental spaces: Aspects of Meaning Construction in Natural Language. Cambridge ; New York, NY, USA: Cambridge University Press.

Manktelow, K. I., & Over, D. E. (1990). Inference and Understanding: A Philosophical and Psychological Perspective. London ; New York: Routlege.

Neural Networks as Fitness Evaluators in Genetic Algorithms: Simulating Human Creativity

Vera Kempe (vera.kempe@stir.ac.uk)
Department of Psychology, University of Stirling
Stirling, FK7 0BQ, UK

Robert Levy (levy@oswego.edu),
Craig Graci (blue@cs.oswego.edu)
Department of Computer Science, SUNY Oswego
Oswego, NY 13126, USA

Introduction

While complex symbolic models of creative processes like music or poetry generation produce remarkable results (Cope, 1996), it may prove advantageous to model creativity more explicitly in terms of adaptive processes of "blind variation and selective retention" (Campbell, 1960)). One computational approach that seems to be particularly promising in this regard is genetic programming (Mitchell, 1996). However, when applying this approach to a complex domain like the creation of works of art, one fundamental problem that arises is the specification of a fitness operator for the selection of the surviving individuals. How can one specify what a good poem or musical piece or painting is? The current model tries to solve this problem by exploring the use of a neural network (NN), which was trained on human evaluations, as fitness evaluator for a genetic algorithm (GA).

Architecture of the Model

The specific domain chosen was limerick generation because of the shortness and clearly defined structure (AABBA) of this poetic form. A lexicon of 1,107 "limerable" words was created using the words of 50 naturally and artificially created limericks. The syllables of each word were represented as binary vectors coding the phonemes and the stress pattern. Information about word class and meaning were left out for the sake of parsimony and computational feasibility. The initial population of limericks was generated by selecting the rhyming words, and then filling each line in accord with the stress template.

Unlike Burton and Vladimirova (1997), who have interfaced an ART NN with a GA for music generation, we were interested in a fitness evaluator that simulates human judgments about limericks. To this end, we obtained quality ratings on a scale from 1 – 6 for 25 naturally and 25 artificially created limericks from 160 participants. A simple recurrent NN was then trained on 36 limericks to associate the median ratings. When tested on the remaining 14 limericks, it produced reliably higher values for natural (3.4) than for artificial limericks (1.9), $t(12) = 2.1$, $p = .05$, as did the human participants (natural: 4.8, artificial: 1.7, $t(12) = 5.4$, $p < .01$). This indicates that the NN clearly captured some important dimensions used by humans to evaluate the quality of limericks. The NN was then interfaced with the GA so as to provide the fitness measure, which was used as the basis for the selection of the fittest individuals from each generation. The selected limericks were then modified using mutation, crossover, and direct-copy operators to create the next generation of limericks.

Simulation Results

The model was run for 1000 generations. In order to determine whether there was improvement in limerick quality, we compared the fitness measures assigned by the NN to the first (1.2) and last 100 (1.6) generations, and observed a small but significant improvement, $t(198) = 9.3$, $p < .001$. This first result encourages us to suggest that training NNs to simulate human judgment in complex domains, and using them as fitness evaluators in GAs, may prove fruitful for the generation of products that typically are dependent on human insight and creativity.

References

Burton, A.R., & Vladimirova, T. (1997) Genetic algorithm utilising neural network fitness evaluation for musical composition. In G.D. Smith, N.C. Steele, & R.F. Albrecht (Eds.), *Proceedings of the 1997 International Conference on Artificial Neural Networks and Genetic Algorithms* (pp. 220-224). Vienna: Springer-Verlag..

Campbell, D. T. (1960). Blind variation and selective retention in creative thought as in other knowledge processes. *Psychological Review, 67,* 380-400.

Cope, D. (1996) *Experiments in Music Intelligence*. Madison, WI: A-R Editions, Inc.

Mitchell, M. (1996) *An introduction to genetic algorithms.* Cambridge: MIT Press.

Modeling the Effect of Category Use on Learning and Representation

Kenneth J. Kurtz (kjk@northwestern.edu)
John T. Cochener (cochener@howard.psych.nwu.edu)
Douglas L. Medin (medin@northwestern.edu)
Department of Psychology, Northwestern University, 2029 Sheridan Rd
Evanston, IL 60208-2710 USA

The formation of categories is based on learning to perform various cognitive tasks and not just categorization per se (see Yamauchi & Markman, 1998). Most models of categorization are committed to explaining the learning, representation, and use of categories solely in terms of taxonomic classification. Recent evidence shows that category structure can be derived from and organized to support the use of examples for particular tasks or goals (e.g., Ross, 1997; Medin, Lynch, Coley, & Atran, 1997).

The ORACLE model (One Representation Across Channels of LEarning) addresses concept formation from the perspective of learning how to represent examples in order to support performance on naturally occurring cognitive tasks. This "you are what you eat" approach stresses that concepts emerge not so much in order to represent the world as it is, but to represent the world relative to the learner's needs and demands. ORACLE is based on two key ideas: 1) input examples are re-represented through error-driven learning to improve task performance (Rumelhart, Hinton, & Williams, 1986); and 2) conceptual organization emerges from numerous iterations of parameter optimization on multiple interwoven processing tasks.

As in a traditional multi-layer network, inputs tend to become represented more closely in the constructed multidimensional space of a hidden layer to the extent they share the same teaching signal at the output layer. ORACLE has two further design principles: 1) different channels of learning lead to different sets of outputs, but share a single set of hidden units; and 2) an anchor channel is dedicated to auto-associative learning (reproducing input information at output). The number and nature of the channels depends upon the tasks or goals the inputs participate in for the learner. Classification is the function being approximated, but the output classes correspond to relevant uses, linguistic labels, or implications; not to established categories. ORACLE predicts that internal representations will tend to emerge that emphasize elements of the input which are most useful for performance across the channels of learning. Therefore, we attempted to simulate Ross' (1997, E1) category use effect.

Participants in Ross' study learned to diagnose fictitious diseases and to select the right treatment. Learning trials were: guess the disease, feedback, guess the treatment, feedback. Each patient consisted of one symptom perfectly predictive for both tasks, one symptom perfectly predictive for diagnosis only, and one non-predictive symptom. After study, participants tested on disease classification performed better on features relevant to both disease and treatment (96%) than on features relevant only to disease (80%). Learning to treat the diseases influenced diagnosis.

The ORACLE architecture consisted of an input layer and a hidden layer of three hidden units (n=3) with projections along three channels of learning to the task outputs (disease, treatment, and auto-association). The layers were feedforward and fully connected. A patient was presented to ORACLE by activating three input units in a 3x12 array based on the twelve possible symptoms and the three possible presentation positions.

Each epoch of standard back-propagation training consisted of randomly ordered presentation of the sixteen patients in all possible symptom orders at a low learning rate. To simulate the two types of training trials, an alternating scheme was used in which half the trials set targets on the disease outputs and half on the treatment outputs. Auto-associative targets stayed set.

The category use effect does not appear initially, but tends to emerge as training proceeds. After 10,000 epochs, 11 of 13 ORACLE runs performed better on features relevant to both tasks. Mean activation of the correct disease output was 96% for the double-relevant and 88% for single-relevance features. ORACLE produces the category use effect by learning anchored representations that support diagnosis and treatment.

References

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology, 32,* 49-96.

Rumelhart, D.E., Hinton, G.E., & Williams, J.R. (1986). Learning internal representations by error propagation. In D.E. Rumelhart & J.L. McClelland (Eds). *Parallel Distributed Processing, Vol. I.* Cambridge, MA: MIT Press

Yamauchi, Y., & Markman, A.B. (1998). Category learning by inference and classification. *Journal of Memory and Language, 39,* 124-148.

Ross, B.H. (1997). The use of categories affects classification. *Journal of Memory and Language, 37,* 240-267.

Towards a Multiple Components Model of Human Memory: A Hippocampal-Cortical Memory Model of Encoding Specificity

Kenneth Kwok (kenkwok@cmu.edu)
James L. McClelland (jlm@cnbc.cmu.edu)
Department of Psychology, Carnegie Mellon University and the
Center for the Neural Basis of Cognition
115 Mellon Institute, 4400 Fifth Avenue,
Pittsburgh, PA 15213, USA

Abstract

While many researchers studying memory today subscribe to a multiple memory systems view of human memory, most existing models of memory do not explicitly embody multiple interacting memory systems in their accounts of human memory performance. The present work is aimed at developing a connectionist model of human memory which: (1) takes into account known principles about the neural basis of human learning and memory; and, (2) is an instantiation of a multiple components memory system along the lines of a *components of processing* framework of memory (Roediger, Buckner, & McDermott, 1999).

The *Hippocampal-Cortical Memory Model* (McClelland, McNaughton & O'Reilly, 1995) comprises two memory components: a hippocampal component which supports rapid learning, and a cortical component which learns more slowly in order to develop integrative representations of the statistical characteristics of the environment across many learning episodes. It provides a suitable starting point for our endeavor because it addresses two important aspects of human memory: (1) the ability of humans to rapidly learn new arbitrary associations; and (2) the fact that memory performance is affected by pre-existing knowledge.

The memory phenomenon that we address here with the model is that of encoding specificity. An early statement of the *encoding specificity principle* (ESP) by Tulving & Thomson (1973) asserts that: *"…specific encoding operations performed on what is perceived determine what is stored, and what is stored determines what retrieval cues are effective in providing access to what is stored"*. This principle was used to explain a pattern of results from a paired-associate memory task in which target words were studied with cues which were weak associates, and memory for the target words was subsequently tested either by cued recall using novel cues which were strong pre-existing associates of the target words, or by recognition. Two key aspects of the results accounted for by the ESP were: (1) the relative ineffectiveness of extralist cues (strong associates of the target words) compared to intralist cues (weak associates) in facilitating recall of the target words; and, (2) recognition failure of words that had been recalled to extralist cues. The crux of this explanation lies in a belief that successful recall and recognition depends on the episodic trace of the event being sufficiently similar to the properties of the retrieval information.

We simulated the ESP experiments by presenting stimuli representing *context*, *relation*, *cue* and *target words* to our model in the form of patterns of activity. After the model had learnt the appropriate associations during a training phase, recall was simulated by presenting the model with cue and context patterns, with the network filling in the target word and relation patterns. Recognition was simulated by presenting the cue and target word patterns, with the network filling in the context and relation patterns.

Our model was able to reproduce Tulving and Thomson's pattern of results in simulations. However, unlike their original explanation which places the burden of successful recall and recognition on an episodic memory system, successful performance of our model in this task relies on both pre-experimental knowledge of word associations in the cortical system and newly learnt associations in the hippocampal system. Hence, in line with a *components of processing* framework, encoding specificity can be understood as a phenomenon that is neither purely episodic nor semantic but relies on both forms of memory.

Furthermore, as a mechanistic instantiation of the ESP, our model is able to make explicit what is meant by statements such as "the trace of the event is sufficiently similar to the properties of the retrieval information" – in a connectionist framework, this is simply pattern completion on a suitable retrieval cue. This in turn provides inroads for further exploration of the parameters and mechanisms which make up a multiple components memory model, and ultimately for characterizing the roles played by different components of the memory system and their interactions, in subserving human memory performance. Further work with the model is being directed at these issues.

References

McClelland J.L., McNaughton B. & O'Reilly R.C. (1995). Why there are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Success and Failures of Connectionist Models of Learning and Memory. *Psy. Rev.* 102, 419-457.

Tulving E., & Thomson D. (1973) Encoding specificity and retrieval processes in episodic memory. *Psy. Rev.* 80(5), 352-373

Roediger,H.L., Buckner,R., & McDermott,K. (1999). Components of Processing. In J. Foster & Jelicic M. (Eds.), Memory: Systems, Process or Function? (pp. 31-65). Oxford University Press.

Categorical Perception as Adaptive Processing of Complex Visuo-spatial Configurations in High-level Basket-ball Players

Eric Laurent (laurent@staps.univ-mrs.fr)
University of the Mediterranean
Faculty of Sport Sciences, av Luminy CP 910
13009 Marseilles, France

Thierry Ripoll (ripoll@newsup.univ-mrs.fr)
University of Provence,
Faculty of Psychology, 29 av Schuman
13621 Aix-en-Provence, France

Hubert Ripoll (ripoll@staps.univ-mrs.fr)
University of the Mediterranean
Faculty of Sport Sciences, av Luminy CP 910
13009 Marseilles, France

Categorical Perception: An Adaptive Process

Research on expertise in sport has provided in the last twenty years explanations about the cognitive processing involved in decision making. The underlying processes of expert performance are known to be related to information stored in memory. Previous studies highlighted the richness of the content of expert conceptual knowledge and the organization of information in memory. This organization related to the chunking hypothesis is thought to reduce the informational load by limiting the number of units that we have to deal with. One hypothesis is that the unitization of elements depends namely on conceptual background of players. Recently, we (Courrieu, Baratgin, Ripoll, Ripoll, & Laurent) tested the assumption that this type of influence could occur in similarity tasks. In several studies, experts were better than novices at discriminating "structured configurations" (*i.e.*, semantically coherent configurations). However some differences (*e.g.*, small physical variations) were poorly detected by experts. Empirical evidence was also given that the influence of knowledge is likely to occur at a perceptually-grounded level. In order to explain this pattern of results we would to test the assumption that knowledge guide perception by modifying perceptual spaces of similarity. Research in the field of categorical perception (*e.g.*, Harnad, 1987) has supported this idea. More recently, Goldstone, Lippa, and Shiffrin (2001) found that the object representations themselves could be altered by category learning. In the same vein we have considered decision making as based on the determination of perception by categorical knowledge.

Building Categorical Material

In order to produce material we used a cluster encoding method validated by Courrieu (2001). Basket-ball coaches built schematic configurations with the following constraint: from one source configuration, drafting two target configurations with an equally physical distortion for both (relatively to the source), but in one case target should belong to the same category as the source and in the other case target should belong to another category.

Categorical Effects

Comparison of novice and expert basket-ball players performances in a same-different judgment task yields evidence for a dissociation of results. Differences between configurations belonging to different categories were particularly well identified by experts while these participants were weaker than novices for detecting differences between configurations belonging to the same category. These data support the idea of a perceptual adaptation constrained by conceptual knowledge and the one of human ability to acquire categorical perception in sport context.

References

Courrieu, P. (2001). Two methods for encoding clusters. *Neural Networks*, *14*, 175-183.

Goldstone, R.L., Lippa, Y., & Shiffrin, R.M. (2001). Altering object representations through category learning. *Cognition*, *78*, 27-43.

Harnad, S. (1987). *Categorical Perception*, Cambridge, Cambridge University Press.

Configural and Elemental Approaches to Causal Learning

M.E. Le Pelley (mel22@hermes.cam.ac.uk)
S. E. Forwood (sef26@hermes.cam.ac.uk)
I.P.L. McLaren (iplm2@cus.cam.ac.uk)
Department of Experimental Psychology; Downing Street
Cambridge, CB2 3EB UK

The nature of stimulus representation in associative networks is a hotly debated issue. On one hand we have elemental theories (e.g. Rescorla & Wagner, 1972) that propose that stimulus compounds are decomposed into their constituent elements, and that learning accrues to representations of these elements. The conditioned responding shown to a stimulus compound is then found by simply adding together the individual associative strengths of each of the elements of that compound. Configural theories (e.g. Pearce 1987) instead posit that a compound stimulus is best viewed as a unitary event separate from its elements, but able to generalise to them. Thus a compound AB is represented by a unit representing the unique configuration "AB". If AB is paired with reinforcement it is this configural unit that develops an association to the outcome (unconditioned stimulus, US). Generalised responding to other stimuli occurs to the extent that these stimuli are similar to experienced configurations.

In our poster we present evidence from a study of human causal learning that bears on this elemental versus configural debate. This study also relates to work on the phenomenon of retrospective revaluation, another research area currently receiving much attention. This involves changes in the strength of previously learned cue-outcome associations in the absence of those cues.

The results of our experiment pose difficulty for some models of human associative learning, particularly those that rely on a configural representation for the cues involved in learning. Taken in conjunction with other work from this laboratory (Le Pelley and McLaren, this issue) that cannot be easily accommodated by elemental theories (e.g. Rescorla & Wagner, 1972), the challenge posed by the data is now sufficiently severe as to require a model employing adaptive parameterisation to govern generalisation (McLaren, 1993, 1994).

The basic design of our experiment is shown in the Table below (other cues were also included so that there were equal numbers of reinforced and nonreinforced single cues and compound cues). Our experiment used an allergy prediction paradigm. This type of paradigm has been used successfully in a number of studies of phenomena of associative learning. However, in order to avoid problems with ceiling effects we adapted this normal allergy prediction paradigm. Thus during training, instead of being asked to judge whether or not an allergic reaction would occur following consumption of a meal, subjects were asked to rate the risk of allergic reaction resulting from these foods (using a four-point scale). Following training, subjects are asked to rate the probability with which a number of individual foods and food compounds will cause allergic reactions. The foods, then, represent the cues, and the allergic reaction is the US. The causal judgment ratings given on test are our index of associative strength.

Stage 1	Stage 2	Test
AB+	A+	A, B, E
CD+	C+	AB, AC, BD,
	EF+	BC, EF, CE&DF

The results are shown below (averaging over equivalent cues). The key findings are (i) retrospective revaluation of cues B&D (backward blocking), (ii) the low rating given to compound BD relative to BC, and the higher rating given to AC, (iii) the high rating of AC relative to EF, and (iv) the fact that CE > BC > DF. These ratings are consistent with revaluation on an elemental basis, but not with configural models employing fixed generalisation coefficients.

References

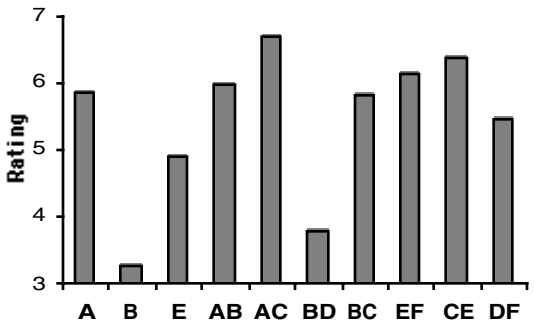
Le Pelley, M.E., & McLaren, I.P.L. (this issue). Representation and generalisation in human causal learning.

McLaren, I. P. L. (1993). APECS: A solution to the sequential learning problem. *Proceedings of the XVth Annual Convention of the Cognitive Science Society* (pp. 717-722). Hillsdale, NJ: Lawrence Erlbaum Associates.

McLaren, I. P. L. (1994). Representation development in associative systems. In J.A. Hogan & J.J. Bolhuis (Eds.), *Causal mechanisms of behavioural development* (pp. 377-402). Cambridge: Cambridge University Press.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review, 94,* 61-73

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A.H. Black & W.F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: ACC



Levels of Processing and Picture Memory: An Eye movement Analysis

Yuh-shiow Lee (psyysl@ccunix.ccu.edu.tw)
Department of Psychology, National Chung-Cheng University
Chiayi, 621, Taiwan, ROC

Introduction

This study examined the underlying mechanisms of the levels-of-processing and picture superiority effects. In contrast to most of the memory research using off-line measurements, this study recorded participant's eye movements during study and testing. Levels-of-processing effects were found not only in verbal materials, but also in pictorial materials such as pictures of faces (e.g., Mueller, Courtois, & Bailis, 1981) and common objects (Lee, 1999). However, using different sets of pictorial materials, Intraub and Nicklos (1985) found the physical superiority effect. They suggest that when visually distinctive pictures are used, having attention drawn to the pictures' unique visual characteristics would lead to a more elaborate trace, than having attention drawn to their semantic attributes. To resolve these inconsistencies, further research examining the role of both semantic distinctiveness and perceptual distinctiveness in producing a better memory for various kinds of pictorial materials is needed. A related issue is that pictures are remembered better than words on conceptually driven memory tests such as free recall and recognition. One of the earliest theories to explain the picture superiority effect is Paivio's dual coding hypothesis. Other researchers focus on the importance of conceptual processing in producing the picture superiority effect. Pictures access meaning codes more directly than words do, and so pictures naturally engage more conceptual processing, which leads to a better memory for pictures than for words.

Eye Movement Analyses

In a typical memory study, participants' differences in encoding or storage of information is inferred from their memory performance in the subsequent testing. In contrast to this off-line measurement, measuring participants' eye movements during study or learning provides an on-line measurement of information processing. Moreover, it has been assumed that the number of features attended to is positively correlated with the number of features actually encoded by participants. Recording the number of eye movements during learning thus makes it possible to observe how many features of a target stimulus were examined. On the other hand, physiological measures can be used to measure the conceptual processing, which is related to the semantic quality of stimuli. In particular, the pupil diameter has been shown to be related to the cognitive processing level in the standard matching paradigm.

Bloom & Mudd (1991) tested *the semantic quality hypothesis* and *the feature quantity hypothesis* in face recognition and found that a deeper processing of face led to an increase in the number of eye movements and an improvement in subsequent recognition performance. Along with evidence from the measure of pupillary dilation, which is considered an index of cognitive effort or processing load, they concluded that the feature quantity hypothesis and not the semantic quality hypothesis was supported. This study investigated whether eye movement analyses can be valid indexes of perceptual distinctiveness/richness and conceptual processing/effort. This question was tested by examining the relationship between eye movement analyses and memory performance.

Results and Discussion

Participant's eye movements during study and testing were recorded. During the study phase, participants studied object pictures, object names, photos of faces, and persons' names. After a distractor task, participants performed a recognition test. The results in memory performance showed that both object pictures and persons' names revealed a ceiling effect. The effects of study condition were significant for object names and photos of faces. Deeper processing produced a better recognition performance. Numbers of eye movements and amount of inspection time were not related to recognition performance, even when the ceiling effect was avoided. Analyses on the pupil sizes also did not reflect the differences in memory performance. In conclusion, recognition accuracy of either verbal or pictorial materials and eye movement behaviors were affected by different variables.

Acknowledgments

This research was supported by the National Science Council of R.O.C. Grant NSC-89-2420-H-194-008.

References

Bloom, L. C., & Mudd, S. A. (1991). Depth of processing approach to face recognition: A test of two theories. *Journal of Experimental Psychology: Learning, Memory and Cognition, 17*, 556-565.

Intraub, H., & Nicklos, S. (1985). Levels of processing and picture memory: The physical superiority effect. *Journal of Experimental Psychology: Learning, Memory and Cognition, 11*, 284-298.

Lee, Y. (1999, November). *Conceptual processing and intentional retrieval in the picture superiority effect.* Paper presented at the 40th Annual Meeting of the Psychonomic Society, Los Angeles, California, USA.

An Alternative Method of Problem Solving: The Goal-Induced Attractor

William B Levy (wbl@virginia.edu)
University of Virginia, Department of Neurosurgery, P.O. Box 800420
Charlottesville, VA 22908-0420 USA

Xiangbao Wu (xw3f@virginia.edu)
University of Virginia, Department of Neurosurgery, P.O. Box 800420
Charlottesville, VA 22908-0420 USA

One theory of problem solving posits solutions by search. That is, the generic problem has a starting point and a goal, where the goal might be precisely known or only sketchily describable. Successful problem solving entails finding a legal (i.e., biologically consistent) path from the starting state to the, or to an, acceptable goal. Influential theories of problem solving, including Newell's work, emphasize the importance of search. That is, the strategy is to try out various paths in hopes that one will lead to the goal. Such searching techniques are computationally intractable in many situations and, in our day-to-day life, we often consider a problem and then find the answer. That is, we find the logical path to the goal without much thought at all and certainly without being consciously aware of trying multiple paths. Here we present a neural network model that solves paradigmatic cognitive problems without search. The alternative to the search strategy in a recurrent neural network is the use of an attractor. An attractor affects the states of a network, and the states of a network are its representations of the world itself. When a network is designed as a sequence-learning network and when it has enough freedom to create its own solutions, i.e., to create novel paths through state space, such a network can find paths from an initial state to a goal state and can find such paths where a path has never before been experienced (Levy, 1996).

The principle of the goal-induced attractor requires or assumes that the system solving the problem has a vague knowledge of the solution. For instance, if I am hungry, I might know I want something to eat, but I might not know exactly where I want to go to eat. We propose that such notions of the goal weakly turn on certain representations. These representations heighten the probability that paths to that goal will be discovered. At the same time, because networks have activity control mechanisms, there will be a tendency not to explore or move towards other goal states. Of course, if the network is not to depend on total randomness, there must be a history of paths learned by the network that can in some way be pieced together by network dynamics.

We use a model of hippocampal region CA3 because this is a sequence-learning region that is capable of coding novel sequences. In particular, and in contrast to error correction-based models, our model is used when mammals do not know the answer and must recode the environment in order to produce simple, usable codes by other brain regions or, from our point of view, by other networks. The problems solved by the CA3 model using the goal-induced attractor are not unlimited but include the simple goal finding problem that is analogous to a rat or a human going from a starting point to a goal by piecing together small paths that have been previously learned but have taken the organism to other places. Other cognitive problems, and some that may even appear to be logical in Nature, can be cast in terms that the goal finding hippocampal model can solve. For instance, the task of transitive inference can be taught to rats, and it can be taught to people in a nonverbal mode. The hippocampal model solves this problem, and it solves the problem, in a sense, by wanting to get the right answer. That is, the goal in performing transitive inference is to get the right answer as opposed to the wrong answer. In this case, the network would have a crude version of the reinforcement "yes, you're right" turned on while it is being presented with the stimuli of the transitive inference task. The task itself is viewed as a sequence but just barely. In particular, the sequence is stimulus, decision/response followed by knowledge of whether the outcome is a success or failure (right or wrong). The model is able to make the right decision for novel, transitive pairings. Another problem that can be solved in a similar manner is the transverse patterning problem.

Our poster will discuss the critical characteristics of our CA3 model that lead to its problem-solving abilities.

References

Levy, W.B (1996). A sequence predicting CA3 is a flexible associator that learns and uses context to solve hippocampal-like tasks. Hippocampus, 6, 579-590.

Sub Space: Describing Distant Psychological Space

Eliza B. Littleton
(littleton@aptima.com)
APTIMA, Inc.,
1030 15th St, N.W., Suite 400
Washington, DC 20005 USA

Christian D. Schunn
(schunn@gmu.edu)
Department of Psychology,
4400 University Drive
Fairfax, VA 22030-4444 USA

Susan S. Kirschenbaum
(kirschenbaumss@npt.nuwc.navy.mil)
Naval Undersea Warfare Center
Code 2214, Building 1171/1
Newport, RI 02841 USA

Submarine navigation proceeds without the benefit of windows. Instead, experienced submarine Officers develop complex skills for mentally turning acoustic information conveyed in displays and in the alphanumerics of passive sonar into spatial representations of other vessels, their paths, intentions, and the high uncertainty of the undersea world (Kirschenbaum, Gray, & Ehret, 1997). We lack data about the psychological space of experts as they reason about distant space in such areas.

Gesture analysis is a promising methodology for observing and understanding the spatial representations that people build to solve difficult problems. It has been used to examine how astronomers reason about the movement of objects when only static representations are visually available (Schunn, Trickett, & Trafton, 1999) and to examine representations of abstractions such as rates of change in algebra word problems (Alibali et al., 1999).

In the current study, we videotaped and analyzed the speech and spontaneous hand gestures of submarine Officers who were instructors in the Submarine Officers Basic Course. Four instructors participated in hour-long, interviews designed to prompt explanations about basic issues in submarining. We transcribed the speech, and then we transcribed the physical aspects of gesture such as the hand or finger shapes, motion, or location in the speaker's personal space.

We found a small set of common gestures that all Officers use similarly (e.g., "bearing" is usually a straight arm/hand, fingers close together pointing not just a direction but a path). We also found a number of topics that Officers convey in gesture.

Our data show that when Officers' gestures convey **perspective** these are more likely to represent an egocentric, as opposed to a removed, exocentric, view. One example is a gesture about sound propagation from the perspective solely of a sensor that is receiving a signal. In other examples, Officers gestured phenomena as relative to themselves physically, as if they were standing in own-ship.

Officers' speech and gesture both convey **uncertainty**, or hedging/estimation. This is important because less experienced Officers are over-certain at times. Often, the only hint that an Officer is hedging or expressing uncertainty will come through one modality only (hands alternate in box-shape up and down; hands circling). About 20 percent of speech and 20 percent of gesture were coded as reflecting uncertainty.

It is common for Officers to gesture complex **spatial dimensions** of the relative positions and motion of the target of interest and own-ship. Officers also gesture complex spatial representations about the shape and degree of angle of conical beams and the effects on signal processing of effects such as the position of the towed array (an array of sensors towed behind the ship) when the ship turns, etc.

Spatial features commonly represented in gesture were shapes (e.g., conical beam, hyperbolic wave front) relative distance and position, direction (bearing, course), and motion (e.g., across or in the line of sight).

Many of the iconic gestures represented or replicated pieces of **common displays** or figures that Officers use. However, other iconic gestures seemed more physical, as if the Officer were relying on an internal, **visual analogy** to describe an idea. In a gesture about beam forming, one hand played the part of the towed array and the other was a vessel being tracked. Each hand moved through the "water," suggesting not only a world-view/bird's-eye view of the contact and own-ship, but the gesture superimposed relative motion on that view, in a way we think is different from other displays and figures.

Acknowledgments

This work was funded by ONR in a contract to APTIMA (# N00014-00-C-0340), a grant to GMU, and support to the Naval Undersea Warfare Center.

References

Alibali, M.W., Bassok, M., Solomon, K.O., Syc, S.E., & Goldin-Meadow, S. (1999). Illuminating mental representations through speech and gesture. *Psychological Science,* 10, 327-333.

Kirschenbaum, S., Gray, W., & Ehret, B. (1997). Subgoaling and subschemas for submariners: Cognitive models of situation assessment. *NUWC-NPT Technical Report* 10, 764-1 (2 June 1997).

Schunn, C., Trickett, S., & Trafton, G. (1999). What gestures reveal about the scientist's mind. Presentation at the Krasnow Institute Lecture Series (November, 1999).

A Criticism of the Conception of Ecological Rationality

Daniel Hsi-wen Liu (hwliu@pu.edu.tw)
Division of Humanities, Providence University
200 Chung-Chi Rd, Shalu, Taichung County 433, Taiwan

The embedded and embodied nature of cognition has been noticed in the 1990s, and has gradually more discussions in philosophy. Discussions centre around the idea that representations are dispensable to a certain extent in the modelling and explanation of cognition (Brooks, 1991; Clark, 1998; Clark & Grush, 1999; Keijzer, 1998; Wheeler and Clark, 1999). Recently, discussions of this nature proceed to explaining the mechanisms of organisms' adaptive flexibility in the ecological niche. The embedded characters of cognition are subtly explored in the notion of 'ecological rationality' (Bullock and Todd, 1999). An interactive-constructive (I-C) approach to modelling intelligence is recently raised, to take into account the dynamical embodied form of adaptiveness (Christensen & Hooker, 2000). This project follows the above trend of discussion but criticises the discussions of organisms' adaptive flexibility.

The primary target of criticism is Christensen & Hooker's (2000) vague account, against which this research will criticise that it begs the question: how is their notion of a 'capacity of coherent, context-sensitive, self-directed management of interaction' carried out on the basis of simple automata? To answer this question will this project argue that the embodied dynamics of cognition is maintained through the recurrent loops of external assessment and internal modification, with a view to manifesting the autopoietic unity of a system's factors, which is originally evident in the maintenance of life. 'Interactive skill construction' is a notion to which Christensen & Hooker (2000) resort in support of the process of 'anticipative skill construction'. At this point they also beg a question: how is self-directed anticipation constructed if no notion of self can be presumed in the cognitive systems? While Christensen & Hooker (2000) see their account as a primary model for cognitive learning, instead will I research in the context of perception, where no obvert functionality of self-control is as evident as learning. With this research will I put their notion of self-directed anticipation in a better profile of explanation.

The explanation envisaged in this research will be cast in terms of stepwise exploitation of environmental information on the basis of inherent a priori representations of the ecological niche. Conceptions that appear in Bullock and Todd (1999) are mainly the domain of decision-making, while I will argue grounded on the domain of perception. Largely against the aforementioned trend of embodied and embedded approach to cognition, but responding to Wheeler and Clark (1999), in this project will I argue for the importance of representations in the embodied and embedded capacities of cognition. On the top of Wheeler and Clark (1999), the previous discussion in this project has provided significant amount of argument, which would in turn bridge a link between representation and the embodied and embedded characters of cognition. With the above argument, this project will criticise Christensen & Hooker (2000) and consequently help the aforementioned trend of embodied and embedded cognition to move ahead.

The main idea of my criticism is that higher-level representations provide guidance in support of low-level organism, while low-level real-time adaptive activities serve to mandate system's processes. Hence the ecological rationality is recurrent between higher and lower level of representations.

References

Brooks, R. A. (1991). Intelligence without representation. *Artificial intelligence*, *47*: 139-160.

Bullock, S., & Todd (1999), Made to Measure: Ecological Rationality in Structured Environments. *Minds and Machines*, ***9***: 497-541

Christensen, W. D., & Hooker, C. A., (2000), An interactivist-constructivist approach to intelligence : self-directed anticipative learning. *Philosophical Psychology*, ***13***(1): 5-45

Clark, A. (1998) Twisted tales: Causal complexity and cognitive scientific explanation. *Minds and Machines* **8:** 79-99**.**

Clark, A., & Grush, R., (1999). Towards a Cognitive Robotics *Adaptive Behavior 7(1): 5-16.*

Keijzer, F. A. (1998) Doing without representations which specify what to do. *Philosophical Psychology*, Vol. 11, No. 3, pp. 269-302.

Wheeler, M. & Clark, A. (1999). "Genic Representation: Reconciling Content and Causal Complexity", *British Journal for the Philosophy of Science*,. ***50***: 103-135.

Thinking through Doing: Manipulative Abduction?

Lorenzo Magnani (lmagnani@cc.gatech.edu)
Department of Philosophy and Computational Philosophy Laboratory, Piazza Botta 6
Pavia, 27100 Italy, and
School of Public Policy and College of Computing, Georgia Institute of Technology, 685 Cherry Street
Atlanta, GA, 30332-0345 USA

Introduction

An interesting and neglected point of contention about human reasoning is whether or not concrete manipulations of external objects influence the generation of hypotheses. I am focusing on the first features of what I call *manipulative abduction* showing how we can find in scientific and everyday reasoning methods of constructivity based on external models.

The Task and the Method

I am analyzing the problem in the light of the so-called historical-cognitive method (Nersessian, 1998). While it tries to integrate findings from research on cognition and findings from historical-epistemological research into models of actual scientific practices, assessments of the fit between cognitive findings and historical-epistemological practices aid in elaborating richer and more realistic models of cognition. There are interesting parallels that can be exploited by cognitive scientists; the relevance of the concept of abduction can contribute to a central issue in cognitive science: hypothesis formation both in science and in everyday reasoning.

Manipulative Abduction in Science

It is well known that many reasoning conclusions that do not proceed in a deductive manner are *abductive*. What I call *theoretical* abduction (Magnani, 2001) is, from a cognitive point of view, an internal process of reasoning. What about the "external" ways of finding hypotheses?

Manipulative abduction happens when we are thinking *through* doing and not only, in a pragmatic sense, about doing. It refers to an extra-theoretical behavior that aims at creating communicable accounts of new experiences to integrate them into previously existing systems of experimental and linguistic (theoretical) practices. Gooding (1990) refers to this kind of concrete manipulative reasoning when he illustrates the role in science of the so-called "construals" that embody tacit inferences in procedures that are often apparatus and machine based.

Epistemic Action

Recent research, taking an ecological approach to the analysis and design of human-machine systems, has shown how expert performers use action in everyday life to create an *external* model of task dynamics that can be used in lieu of an internal model (Kirlik, 1998). Not only a way for moving the world to desirable states, action performs an *epistemic* and not merely performatory role that is very relevant to abductive reasoning.

Experiments and the "World of Paper"

Already in the *Dialogues Concerning the Two Chief World Systems* (1632), accentuating the role of observational manipulations Galileo presents an anatomist that, manipulating a cadaver, is able to get new, not speculative, information that goes beyond the "world of paper" of the Peripatetic philosophy. It is well known that recent philosophy of science has paid a great attention to the so-called theory-ladenness of scientific facts (Hanson, Popper, Kuhn). Nevertheless a lot of new information in science is reached by observations and experiments, and experiments are the fruit of various kinds of artifactual manipulations: the different strategies correspond to the expert manipulations of objects in a highly constrained experimental environment, directed by *abductive* movements that imply the application of old and new extra-theoretical *templates* of behavior.

What I Expect to Find

We still know very little about what governs the manipulative abduction. I plan to better delineating some of the manipulative *templates* of behavior that are active in creative abduction comparing scientific and everyday reasoning: 1. *simplification* of the reasoning task; 2. capacity of overcoming situations of *incomplete* or *inconsistent* information; 3. *control of sense data*: we can change the position of our body (and/or of the external objects) and exploit various kinds of prostheses (instruments, etc.); 4. *external "models"* of task mechanisms.

References

Gooding, D. (1990). *Experiment and the Making of Meaning.* Dordrecht: Kluwer.

Kirlik, A. (1998). The ecological expert: Acting to create information to guide action. *Proceedings of the 1998 Conference On Human Interaction with Complex Systems*, (HICS'98). Piscataway, NJ, IEEE Press.

Magnani, L. (2001). *Abduction, Reason, and Science. Processes of Discovery and Explanation.* New York: Kluwer Academic/Plenum Publishers.

Nersessian, N. J. (1998), Kuhn and the cognitive revolution. *Configurations, 6,* 87-120.

Spatial Priming of Recognition in Virtual Space

Gareth E. Miles (MilesGE@cf.ac.uk)
School of Psychology, University of Wales, Cardiff, P.O. Box 901, CF1 3YG, UK.

Andrew Howes (HowesA@cf.ac.uk)
School of Psychology, University of Wales, Cardiff, P.O. Box 901, CF1 3YG, UK.

Introduction

Virtual environments are often not veridical facsimiles of reality. Efficiencies of navigation are often made available to users with the use of hyperlinks but other schemes that violate the normal rules of Euclidean space are also possible (Ruddle, Howes, Payne & Jones, 2000). A virtual environment user may experience hundreds of different locations all with the same apparent Euclidean co-ordinates if space is allowed to overlap itself. It is anticipated that methods for probing subjects spatial representation that involve strategic processes (such as distance estimation and map drawing) will be poor for evaluating how subjects represent the discontinuities and spatial overlaps that occur in this kind of space. In most cases the unusual features of the space will be highly salient and are likely to input into any strategic process, distorting evidence about the representation of spatial information. An important advantage of a priming methodology in this case is the absence of any strategic processes – priming data are claimed to derive directly from the underlying representation of a stimuli. One aim of our current work is to establish the validity of priming methodologies for revealing human representation of virtual environments. We hope to achieve this by replicating work done using 2D map representations.

McNamara, Halpin & Hardy (1992) used priming in item recognition and location judgement to assess the relative contributions of order of presentation (temporal proximity) and spatial proximity on their participants representation of a two dimensional map. The experiment summarised in this paper used a similar design, however, instead of using a two dimensional map with object locations represented as dots we use a 3-dimensional virtual environment with object locations represented by small virtual cubes.

Experiment

Thirty-two participants navigated an experimental environment. Participants used the mouse to control where they looked in the 3D environment and the space bar to 'walk' through the environment in the direction faced. The experimental environment consisted of a large 'warehouse-like' triangular room, in which twelve items were located, and an antechamber from where subjects began. Subjects completed a training phase in which they were shown the location of the twelve items and then completed a test phase during which they had to indicate the correct location of each item. Training and test phases were iterated until the subject had successfully remembered the location of all the items. The twelve items were divided into four filler items, and four sets of pairs. Each pair was assigned to one of the experimental conditions (a 2x2 design with spatial: close-distant, and temporal: close-distant).

After the experiment, subjects were given recognition and then location judgement tasks. Immediately following a warm-up task the subjects were told that in the next section they had to decide whether the named items were included in the 3D environment they had learned. The twelve items from the 3D environment and twelve foils made up the list of item names presented. The paired items were presented consecutively, with one item, in each pair, acting as a prime and one acting as the target. These 24 item names were presented three times in the same order.

Results and discussion

A repeated measures ANOVA on the recognition response time data found a main effect of spatial proximity, $F(1, 31) = 4.62$, $MSE = 6980$, $p < .05$, no effect of temporal proximity, $F(1, 31) = 2.00$, and no interaction of spatial and temporal proximity, $F(1, 31) = .309$.

We are using these results to inform the building of computational models that learns the locations of the objects in the virtual environment. Spatial priming at close spatial and temporal proximity can be explained by a model that encodes each item's heading from the antechamber and any errors that are made whilst trying to find that item. However, to account for the main effect of spatial proximity a model using either metric or propositional information about relative object locations is required.

References

McNamara, T. P., Halpin, J. A., & Hardy, J. K. (1992). Spatial and temporal contributions to the structure of spatial memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 18,* 555-564.

Ruddle, R. A., Howes, A, Payne, S. J., & Jones, D. M. (2000). The effects of hyperlinks on navigation in virtual environments. *International Journal of Human-Computer Studies, 53,* 551-581.

The frequency of connectives in preschool children's language environment

Bradley J. Morris (bmorris@andrew.cmu.edu)
Dept. of Psychology, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213

The basis of logical reasoning is a set of logical connectives such as AND, OR, and NOT. The rate at which each is produced and comprehended by young children differs significantly between connectives. For example, on average children produce AND and NOT around 27 months, yet do not produce OR until age 5 (Fenson, et al, 1994). One possible explanation for these differences is the frequency of logical connectives in children's language environment. The little research jointly examining children's logical reasoning and their language environment (e.g., Scholnick & Wing, 1992) suggests that children's language environment does influence the rate of understanding and production of logical connectives.

The present study examined transcriptions of audiotaped conversations obtained through the CHILDES database (MacWhinney, 2000) to compare the frequency of production and reception of connectives by children between the ages of 2.0-5.5 in conversational situations with at least one adult. The connectives selected were AND, OR, NO/NOT (NOT was also coded for the number of uses as a word and in the form of a contraction). A total of 187 databases were coded containing 81,676 conversational turns (mean = 456.2, SD = 282.6).

Results and Discussion

A total of 10,201 No/Nots, 5,149 Ands, and 396 Ors were produced. Table 1 displays the distribution of connectives per 100 conversational turns by producer (adult/child) and age of the child (six-month intervals based on the child's age during data collection).

Overall, NOT was more frequent than AND and OR (Fs (2, 165) = 4.9 and 22.7 respectively, ps <.01), while AND was more frequent than OR (F (2, 165) = 10.6, p<.01). Results suggest significant differences in the use of NOT across age groups (F (6, 102) = 3.7, p<.002) but no significant differences for AND and OR (p>.10).

When split by producer (adult v. child), adults produced OR and AND significantly more than children (Fs (6, 102) = 4.3 and 3.2 respectively, ps <.05 and significantly more negations as contractions than non-contractions (F (6, 102) = 3.7, p<.002).

Age-related changes for children's connective use indicate significant increases in the use of OR (F (6, 123) = 5.2, p<.001), AND (F (6, 123) = 3, p<.01), and the contraction form of NOT (F (6, 123) = 4.1, p<.001) from age 2 to 5.5. Changes in adult usage by child's age were significant only for NOT as a contraction (F (6, 132) = 3.4, p<.01).

Table 1- Connectives per 100 conversational turns

	NOT (Contractions)		AND		OR	
Age	Adult	Child	Adult	Child	Adult	Child
2-2.5	2.63 (2.80)	2.23 (0.53)	2.98	0.98	0.26	0.01
2.5-3	2.06 (3.69)	2.82 (1.68)	3.23	2.97	0.40	0.02
3-3.5	2.60 (3.98)	3.06 (2.47)	3.97	2.59	0.38	0.05
3.5-4	2.93 (4.38)	3.45 (3.21)	3.39	3.00	0.37	0.19
4-4.5	2.99 (3.97)	3.08 (3.16)	3.32	2.99	0.45	0.21
4.5-5	3.00 (5.02)	3.04 (3.58)	3.63	3.95	0.46	0.11
5-5.5	3.47 (2.62)	2.70 (3.45)	3.79	2.86	0.32	0.14
Mean	2.73 (3.81)	2.90 (2.38)	3.46	2.62	0.37	0.09

Overall, the most notable difference is the overall production levels of each connective. NO/NOT is approximately 25.7x as frequent as OR and twice as frequent as AND while AND is 13x as frequent as OR in conversation. Children's use of contractions (e.g., don't) increases significantly with age. AND use by children increases dramatically from ages 2-4 to a level by age 3.5 that is similar to the levels adults use when talking to children. The use of OR by both children and adults is significantly lower than AND and NOT and children's production rates remain well below the levels adults use when talking to children. Thus, results suggest that those connectives produced at earlier ages (e.g., NOT) are associated with higher levels of adult production than connectives produced at later ages. One explanation for these results is that adult production rates may vary in proportion to child comprehension rates exhibiting a reciprocally causal relationship.

References

Fenson, L., Dale, P.S., Reznick, J. S. , Bates, E., Thal, D.J., & Pethick, S.J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development,58* (5-Serial No. 242).

MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk, Vol 1: Transcription format and programs (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Scholnick, E.K., & Wing, C.S. (1992). Speaking deductively: Using conversation to trace the origins of conditional thought in children**.** *Merrill-Palmer Quarterly*, 38(1), 1-20.

A Soar model of human video-game players

Hidemi Ogasawara (hidemi@sccs.chukyo-u.ac.jp)
School of Computer and Cognitive Science, Chukyo University; 101 Tokodate, Kaidu-cho
Toyota, Aichi 470-0393 Japan

Takehiko Ohno (takehiko@brl.ntt.co.jp)
Communication Science Laboratories, NTT; 3-1 Morinosato, Wakamiya
Atsugi, Kanagawa 243-0198 Japan

Introduction

The real world environment around humans is mostly dynamic. Choosing a video-game "Pac-Man" as a representative of the dynamic situation typical of everyday life, we have explored players' behavior through psychological experiments and a case study (Ogasawara & Ohno, 1999). "Pac-Man" is a game where a player controls a character called Pac-Man to eat dots on a maze while escaping from four enemies called ghosts which chase Pac-Man. Also, under some specific conditions, Pac-Man can kill ghosts to get extra points.

In this presentation, we discuss a computational model of the player implemented in the Soar architecture. To model the real time task performance, we extended the Soar architecture by adding perceptual and motor processors, and by synchronizing both the game progress and I/O processor with the cycles of the model. In the following section, we explain our Soar extensions and a player model on the extended Soar architecture.

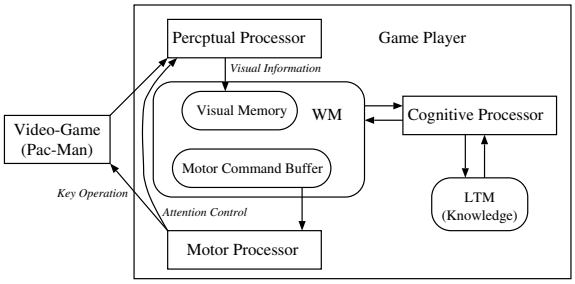


Figure 1: Diagram of the Model

Model

The Soar-8 architecture is the base of our model. The perceptual (input) and motor (output) processors are added to interact with the game (Fig. 1). The basic concept of the architecture is similar to another extended Soar architecture, NOVA (Wiesmeyer, 1992), with a few differences in visual information representation and a synchronization method.

The perceptual processor generates visual information in the working memory (WM). The visual information of the game is represented by relations between static objects (e.g., "road", "corner", etc...), dynamic objects ("Pac-Man", "ghost", etc...), and their state. Some meta-representations like "sequence of dots" are also included in the visual information. The quality of visual information differs between the attended area and the unattended area. This attention mechanism is modeled by controlling "focus" through the motor processor, and by obscuring some relations, like connection information between "roads" and "corners", in the peripheral area.

Synchronization between the game and the model is based on the "decision cycle" of Soar, which is estimated about 100 ms (Newell, 1990). Also, the delay for perceptual and motor processing is implemented by delaying transfer between WM and the processors for one decision cycle.

We test this extended architecture with a player model that uses keystroke level operators and attention control operators directly. The model is based on an "escaping from enemies" strategy that is effective for many action-type video-games. This naive model "looks at" Pac-Man, and attempts to chase dots and avoid ghosts. Our human player data shows that novices mainly depend on this strategy and look at Pac-Man more often than experts do. Though we have not examined its behavior with the human players' data closely, this model earns about the same points as novices do. We will construct another model grounded on our extended GOMS analysis (Ohno & Ogasawara, 1999) of a human player's data.

References

Newell, A. (1990). *Unified Theories of Cognition*. Harvard University Press, Cambridge, Massachusetts.

Ogasawara, H.,& Ohno, T. (1999). Expertise process of human video-game players. In *Proceedings of ICCS/JCSS99*, pages 388–393.

Ohno, T., & Ogasawara, H. (1999). Information acquisition model of the highly interactive tasks. In *Proceedings of ICCS/JCSS99*, pages 288–293.

Wiesmeyer, M. D. (1992). *An Operator-based Model of Human Covert visual Attention*. PhD thesis, University of Michigan.

Practical Cognition in the Assessment of Goals

Luis A. Pérez-Miranda (ylppermil@sf.ehu.es)
Institute for Logic, Cognition, Language, and Information (ILCLI, UPV-EHU)
P.O. Box 220, 20080 Donostia-San Sebastian (Spain)

1. Practical Cognition

Practical cognition seems to help the agent in the way of constructing strategies and plans in his pursuit of a better situation for him. The goals and objectives of an agent can be of diverse nature, from mere intrinsic desires to sub-goals of already pretended plans. Practical cognition can be seen as the basis on which the process of selecting and executing courses of actions for achieving those goals (using plans or operators) takes place. One of the tasks of practical cognition is to cope with conflict situations of decision-making among agent's potential goals. Because agents have incomplete knowledge about the world, it is inevitable that some of these goals will conflict (Ferguson, 1992; Pryor, 1994). Sometimes an agent is forced to choose among different relevant options that are jointly incompatible to pursue.

Our approach assumes that, not always, but in many cases, the adoption of goals is plan dependent. Generally, it happens that a goal cannot be adopted before the agent realizes that is able to bring a plan about for the occasion. Often an important amount of the value of a goal is directly obtained from the expected utility value of the plan where it is embedded (Beaudoin, 1994). The adoption of a goal is related to three factors: the value of the goal itself, the possibility of constructing a plan pursuing a previously learnt strategy for that goal, and agent's commitments related to previous plans (Pérez Miranda, 1997).

Once the agent has recognized that a potential goal is obtainable, the next step in determining the adoption of a goal is to detect any adverse effects between that goal and other likely pretended goals derived from previous intended plans or single urgencies that ought to be accomplished without delay. Hence the agent must look for scenarios in which both potential goals and ongoing adopted goals fit together insofar as fulfilling one may be at odds with fulfilling another or with maximum fulfillment of the overall set. We are concerned with explaining how an agent could arrive to manage and fit these factors in a suitable way as to behave, so to speak, following some rational patterns.

2. The Filtering Mechanism

The evaluative mechanism proposed here only concerns with those goals that have a motivational or cognitive grounding (or both together). Beliefs are the unique available evidence for an agent to make decisions about whether what he wants to do is or not justified under the circumstances. We think this connection between beliefs (or motivations) and goals can be encoded into an ordered pair, the reason supporting the goal, and be evaluated according to order and strength criteria.

Order among supporting reasons constraints the decision process to only those decisions that are relevant for the agent while just excluding or postponing the others. In particular, high order reasons override low order reasons, ruling them out from the process of assessment. Furthermore, ordering reasons is a way of facing situations of apparent incomparability, for instance, among supporting reasons that are desires and reasons that are beliefs. Strength determines the expected degree of utility derived from adopting or not a goal at a point time given the evidence available.

Our filtering mechanism selects only those goals whose supporting reasons result undefeated according to agent's doxastic states. The mechanism embodies two levels of decision-making attending to the order and strength of the supporting reasons. An agent only would be justified in adopting a goal when the reason that supports that goal results undefeated.

Acknowledgments

This work has been supported by the Research Project 1/UPV/EHU 00109-HA-4487/1998.

References

Pérez Miranda, L.A. (1997). Deciding, Planning, and Practical Reasoning: Elements Towards a Cognitive Architecture. Argumentation, 11, 435-461.

Pryor, L. (1994). Opportunities and Planning in an Unpredictable World. M Phil Thesis (Computer Science), Northwestern University, Evanston, Illinois.

Beaudoin, L. P. (1994). Goal Processing in Autonomous Agents., PhD thesis, School of Computer Science, The University of Birmingham.

Ferguson, I. A. (1992). Toward an Architecture for Adaptive, Rational, Mobile Agents. In Werner, E. and Y. Demazeau (Eds.) (1991). Decentralized AI 3. Proceedings of the Third European Workshop on Modelling Autonomous Agents in a Multi-Agent World. Amsterdam: Elsevier.

Exceptional and temporal effects in counterfactual thinking

Susana Segura (s_segura@uma.es)
University of Malaga. Facultad de Psicologia.Campus de Teatinos, s/n 29071 Malaga. Spain.
Rachel McCloy (rmccloy@tcd.ie)
University of Dublin. Department of Psychology. Trinity College. Dublin 2. Ireland.

Introduction

Counterfactual thinking is the comparison of a factual situation to a simulated alternative situation. When people think about how things could have been different, they tend to change some kinds of factual events more often than others, for example, people have been shown to be more likely to undo the *last* event in an independent sequence (the *temporal order* effect; Byrne, Segura, Culhane, Tasso, Berrocal, 2000; Miller and Gunasegaram, 1990). Research has to date focussed on separating out each of the different factors that determine the events which people choose to change. We know little about how such factors interact.

McCloy and Byrne (2000) showed that people are more likely to change inappropriate controllable events, which are exceptional with regards to interpersonal norms for behavior, than to change appropriate events, which are normal. The first aim of this study was to establish whether varying interpersonal normality on a different dimension (selfishness) could affect the mutability of events. We predicted that controllable events which were *selfish* (i.e., done purely for the self) would be more mutable than controllable events which were *selfless* (i.e., done for more than just the self), because selfish events are exceptional with respect to interpersonal norms and selfless events are normal. The second aim of this study was to explore the interaction between interpersonal normality and the temporal order effect described above. We predicted that the selfishness of an event would interact with its position in a temporal sequence, in that selfish events which occur last in an independent temporal sequence should be changed more often than selfish events which occur earlier.

Method

We constructed a scenario describing a morning in the life of a woman (Maria). On this day, Maria carried out nine tasks before leaving the house. Of these, eight were everyday household tasks (e.g., dusting, ironing) while one was an event carried out purely for Maria herself (reading a magazine). On leaving the house she had a car accident. Participants were asked to imagine how the outcome might have been different if she had not done one of the tasks.

The within-subjects variable was the nature of the tasks (selfish vs. selfless). The between-subjects variable was the position in which the selfish event appeared in the scenario. The selfish event could appear as either the first event, the fifth (middle) event, or the ninth (last) event in the scenario. The dependent variable was the task that participants chose to undo. The participants were 117 undergraduates from the University of Malaga, Spain. They were assigned at random to one of the three experimental conditions regarding the position of the selfish event (first, n = 35; middle, n = 42; last, n = 40).

Results and Discussion

The results corroborated both of our hypotheses. Overall, the selfish event (30%) was more mutable than any of the selfless events (all <9%). These results provide a replication of the effects of *interpersonal normality*, and they extend this effect to events that deviate from prevailing interpersonal norms along another dimension (selfishness). As predicted, the results show that there is an interaction between the effects of interpersonal normality and temporal order. The selfish event was more mutable when it occurred in the last position (40%) than when it occurred in the first (31%) or middle (19%) positions. However, the results do show that the selfish event is still highly mutable in the first position. One explanation may be that the occurrence of the selfish event early in the scenario may violate people's assumptions about when people normally carry out such actions. A selfish action carried out when there are selfless waiting to be done may be seen as more exceptional than one carried out after the completion of such tasks. Nevertheless, the results of this experiment do show that selfish event are more mutable than selfless events, and that the effects of interpersonal normality and temporal order can interact in determining counterfactual mutability.

References

Byrne, R.M.J., Segura, S., Culhane, R., Tasso, A., and Berrocal, P. (2000). The temporality effect in counterfactual thinking about what might have been. *Memory and Cognition, 28,* 264-281.

McCloy, R. and Byrne, R.M.J. (2000). Counterfactual thinking about controllable events. *Memory and Cognition, 28(6),* 1071-1078.

Miller, D.T. and Gunasegaram, S. (1990). Temporal order and the perceived mutability of events: implications for blame assignment. *Journal of Personality and Social Psychology*, *59*, 1111-1118.

Children's Algorithmic Sense-making through Verbalization

Hajime Shirouzu (shirouzu@sccs.chukyo-u.ac.jp)
School of Computer and Cognitive Sciences, Chukyo University
Toyota, Aichi 470-0393 JAPAN

Introduction

This paper demonstrates the effectiveness of children's own verbalization on their conceptual understanding of why they do what they do to solve a simple arithmetic problem. The problem was solvable by the interaction with the external resources, and the externalized answers could be described verbally as they were seen. Verbalization, however, in its essence, could include talker's own interpretation or explanation of the externalized records (Pine & Messer, 2000; Shirouzu, Miyake & Masukawa, 2001).

I conducted a small-case learning experiment, asking six sixth-graders in a class to cut out the 3/4 of 2/3 of the origami paper's area. They were of roughly same performance on the math and had already mastered the fractional multiplication. Initially, all of them manipulated the paper directly to solve the task. Yet, gradually guided by a teacher-experimenter, through multiple collections of the solutions and explicit comparisons among them, four students actively worked out why the answer was equal to the one-half of the whole and finally verbalized its algorithmic solution (2/3X3/4=1/2). Six months later, these students described the task by mentioning its algorithmic aspect as "devising various ways to make the one-half area," but the remaining two could not do so even though they also gave explicit consent to the algorithmic solution proposed at the end of the lesson. I hypothesize that the key to the individual differences is in their verbalization on how they interpreted own externalized solutions, differences or similarities among peers' solutions, and the task itself.

Learning Setting

The data come from a 6th grade classroom in a remote branch school of Japan, which had six students (2 girls, 4 boys) as a whole. I visited there twice as a teacher-experimenter to conduct a lesson and make a follow-up inquiry, both of which were recorded by videotape for analyses.

There were three intended phases in the lesson, to make students solve the problem and explain their solving steps, to have them reflect upon the differences or similarities among solutions, and to ask them why the goal area was constant as one-half. For the first phase, I prepared sheets of origami paper, a pencil, and a pair of scissors, and then asked them a problem. Every time the student presented his answer, I accepted it and made him explain to all how he made it with visualizing solving processes by extra origami papers. For the second phase, I let the students compare each two solutions chosen from what they had made. For the third phase, I asked what was common among all answers and why it was.

Six months later from the lesson, I visited the class again with the inquiry: "Please write down anything that you remember about what happened at the last lesson."

Analysis

Overall, the performance of this class "appears" to be quite high. Everyone solved the given task actively and correctly. Newer interpretations were frequently made and easily shared under "one voice." Hidden by such seemingly one voice, however, crucial differences in their understanding occurred through chances of verbalizing their own interpretations.

If a student replied to the question about the sameness of the answers as, "They are the same not in form but in area," instead of only as, "Different," I coded that he verbalized more than what was seen actually. When the others only consented to such interpretation, I defined that they did not take initiative of explicit verbalization. In this way I coded what child mentioned what interpretation. Although space prevents me from describing the entire shifting-process of interpretation, the interpretations they made and articulated in the lesson appeared in their reports in the follow-up inquiry clearly.

Child 1, for example, answered to why all the solutions were the same as, "If I multiply these two fractions, we can see the answer in the frame of the whole, which equals one-half. So, all of these are equal to the one-half of original." In the following-up, he tied his experience of using origami to the fractional multiplication. On the other hand, Child 2 consented vigorously to Child 1's explanation above, but answered to the inquiry, "I remembered two-thirds and three-fourths," which reveals his remembrance of fragmental facts. Child 3 explained her solving step of the second try as, "If a part of the rest (1/4 of 2/3) of my first answer is combined with this area (1/3), I can get three folded rectangles. This (the answer, 3/4 of 2/3) has also the three rectangles. So if I folded the paper into the half of six parts, the three-sixths, I thought that I could make the 3/4 of 2/3." Even though this early reference to the one-half-ness could not be shared among others, she was explainable at the end of the lesson why the answer was one-half based on her diagrammatic understanding, which could be also recognized in her following-up report.

The result implies that the students who verbalized their interpretations could produce durable abstract understanding. This proposes a protest to the lecture style instruction in which a teacher delivered well-structured explanations and students are only silent. Instead, we have to make careful analyses on each student's talk and trigger finer interactions to promote the externalization of their own explanations, ultimately to let them deepen their learning by themselves.

References

Pine, K. J., & Messer, D. J. (2000). The effect of explaining another's actions on children's implicit theories of balance. Cognition and Instruction, 18, 35-51.

Shirouzu, H., Miyake, N., & Masukawa, H. (2001). Cognitively active externalization for situated reflection. Submitted to Cognitive Science

Prosodic Guidance: Evidence for the Early Use of Capricious Parsing Constraint

Jesse Snedeker (jessned@psych.upenn.edu),
John Trueswell (trueswel@ psych.upenn.edu)
Institute for Research in Cognitive Science; 3401 Walnut Street, Suite 400A
Philadelphia, PA 19104 USA

Abstract

This experiment examines the use of naturally occurring prosodic cues to syntactic structure. Earlier work suggests that the production of informative prosodic cues depends upon speakers' knowledge of the situation: speakers provide prosodic cues when needed; listeners use these prosodic cues when present (Snedeker et al., 2001). The present experiment explores the online use of prosodic information in a world-situated eyegaze task. A referential communication paradigm was used to elicit productions of ambiguous sentences (*Tap the frog with the flower*) and determine whether listeners could use prosodic cues to correctly interpret them. Acoustic analyses indicated that Speakers produced potentially informative prosodic cues. Listeners' responses to the ambiguous sentences strongly reflected the demonstration the Speaker had seen, indicating that they were able to use this information. Analyses of the eye-movements indicate that listeners use prosodic information to inform their parse shortly after the onset of the direct-object noun.

Introduction

Prior work on the use of prosody in online sentence processing suggests that strong cues can have an early influence on parsing preferences (for review see, Kjelgaard & Speer, 1999). Prosody in these studies was typically manipulated by splicing pauses into speech to indicate clause boundaries, manipulating synthesized speech, or asking trained speakers to produce particular prosodic variants of an utterance. But in naturally occurring speech, syntactic structure is only a weak predictor of prosodic variation (for review see Fernald & McRoberts, 1996) Unsurprisingly, a number of researchers have found that naïve speakers produce less consistent prosodic cues for syntactic disambiguation than the informed speakers typically used in comprehension experiments (e.g., Allbritton, McKoon, & Ratcliffe, 1996) The research reported here examines whether listeners' use the prosodic cues produced by naïve speakers in real time.

Methods

This experiment employed a referential communication task, in which a Speaker and a Listener were separated by a divider, allowing for only verbal communication between the two participants. Both participants were given identical sets of objects. The Speaker delivered memorized directions, in an attempt to get the Listener perform particular actions on the set of objects on her side of the screen.

The critical sentences contained globally ambiguous prepositional phrase attachments, such as "Tap the frog with the flower". The phrase "with the flower" can be taken as Instrument (VP-Attachment) or a Modifier (NP-attachment). On each trial both of the subjects received: 1) a Target Instrument, a full scale object that could be used to carry out the action (e.g., a large flower); 2) a Marked Animal, a animal carrying a small replica of the instrument (e.g., a frog holding a little flower); 3) an Unmarked Animal (e.g., an empty-handed frog); and 4) two unrelated objects. The set of toys supported both interpretations of the ambiguous sentence. The Experimenter demonstrated one of two possible actions: an Instrument action (e.g., the Experimenter picked up the large flower and tapped the plain frog) or a Modifier action (e.g., using her hand, the Experimenter tapped the frog that had the small flower).

The Listener wore a head-mounted eyetracker. Trained coders viewed the videotape of the Listener's eye-movements and recorded the onset of each word in the target sentence and the onset and location of each fixation that occurred from the beginning of the instruction until the subject began the action.

Results and Discussion

Demonstration Type (Instrument or Modifier) had a reliable effect on the Listener's actions, the length of the post-verbal pause, the length of the direct-object noun and the pause that followed it, and the length of the prepositional phrase. Demonstration Type also had an early effect on Listener's eye-movements. Within 400 ms after the onset of the direct object noun, Listeners in the Modifier condition showed a preference for the Marked Animal, while Listener's in the Instrument condition divided their fixations between the Marked and Unmarked Animals. We conclude that prosody can be used to rapidly resolve an attachment ambiguity.

Acknowledgments

We thank Michael Felberbaum and Nicora Placa for their assistance, input and patience. This work was supported by NIH Grant 1-R01-HD3750707-01 and a NSF Center Grant to the Institute for Research in Cognitive Science.

References

Albritton, D., McKoon, G. & Ratcliff, R. (1996). Reliability of prosodic cues for resolving syntactic smbiguity. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **22**, 714-735.

Fernald, A. & McRoberts, G. (1996). Prosodic bootstrapping: A critical analysis of the argument and the evidence. In J. Morgan & K. Demuth (Eds), *Signal to Syntax* Mahwah, NJ: Erlbaum.

Kjelgaard. M. & Speer, S. (1999). Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *JML*, **40**, 153-194.

Learning and Memory: A Cognitive Approach About The Role of Memory in Text Comprehension

Soares, Adriana Benevides Soares (adriana@uenf.br)
Universidade Estadual do Norte Fluminense, Brazil
Corrêa, Carla Patrícia Quintanilha (cpcorrea@zipmail.com.br)
Universidade Estadual do Norte Fluminense, Brazil

This work tries to understand the organization of knowledge in memory. There are two theories about the mnemonic system. One of them says that memory is formed by two distinct memory stores: a short-term memory, and a long-term memory. The other theory involves semantic activation models to explain memory working. We have carried out three experiments to investigate this subject. In these experiments, we have employed a probe technique to observe the semantic representation in memory during text comprehension. Our results showed response times increasing as a function of time, and that seems to be consistent with models involving semantic activation. This conclusion supports the study of learning and comprehension texts. Our goal is to offer a learning strategy which would improve text comprehension. This strategy, based on the work of Yekovich and Walker (1986), allows the possibility of building a text in order to improve the process of comprehension. This improvement ocurrs through the activation of informations by peripheral concepts. An experiment has been conducted to verify this strategy. We have built two different texts to observe which one would be better understood. The results showed that the text built through our strategy was indeed better understood. Therefore, we argue that using peripheral concepts is relevant to the process of comprehension, because it activates the text's central concept, improving the comprehension's process.

References:

Achour, L. & Le Ny, J. F. (1983). L'évolution de la représentation évoquée par un mot au cours de la compréhension de phrases: Étude par la technique du sondage. L'Année Psychologique, 83, 409-422.

Hyodo, M., Le Ny, J. F. e Achour, L. (1994). The corse of representation in memory during the compreension of paragraphs. Année Psychologique, 29(5), 565-590.

Yekovich, F. R. e Walker, C. H. (1986). *Retrieval of Scripts Concepts* Journal of Memory and Language 25, 627-644.

SARAH: Modeling the Results of Spiegel and McLaren (2001).

Rainer Spiegel (rs272@cam.ac.uk)
University of Cambridge, Department of Experimental Psychology, Downing Site
Cambridge, CB2 3EB, UK

IPL McLaren (iplm2@cus.cam.ac.uk)
University of Cambridge, Department of Experimental Psychology, Downing Site
Cambridge, CB2 3EB, UK

This hybrid associative/cognitive model simulates the results of our other paper at this conference, in which we demonstrated that a particular sequence learning task cannot be explained by the entirely associative SRN alone. Our model adds cognitive mechanisms to the SRN, for which we had found evidence in the structured interviews given to subjects (2001a,c). Subjects had verbalized repetitions of circle flashes, symmetries within each sequence and that each sequence had the basic structure ABCBA. The interview also revealed that they made analogies between the sequences they had experienced in training, and we hypothesized that this may have helped them to generalize to the novel sequences on which they were tested. Each cognitive mechanism was implemented as an autonomous agent and was assigned a certain probability of being carried out, e.g. 14.29 percent of the subjects verbalized symmetries, hence p=.1429. The other probabilities can be found in (2001c). More detailed explanations of SARAH (Sequential Adaptive Recurrent Analogy Hacker) and a related hybrid model can be found in Spiegel and McLaren (2001b,c).

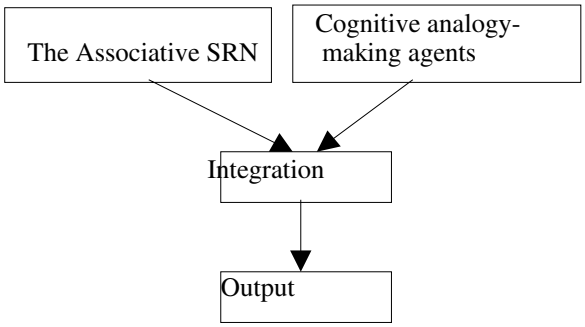


Figure 1: The SARAH model architecture

Training the same number of SARAHs (30) as there were people in the experiment (2001a) on entirely the same task, with six hidden units, a learning rate of .1, 40000 training trials and carrying out the same kind of ANOVA as on the empirical data revealed a significant main effect for the between subjects factor *group*, $F(1,28)=52.78$, $p<.001$, $f=1.37$, $\eta^2=.65$. The Experimental SARAHs ($M_e=.43$, $\pm SE_e=.05$) reveal a significantly higher activity difference when compared with the Control SARAHs ($M_c=.09$, $\pm SE_c=.01$). When considering the novel sequence type (2 Cs), the results resemble the human subjects (2001a) in the way that the SARAHs in the Experimental group show significantly better generalization when compared with the SARAHs

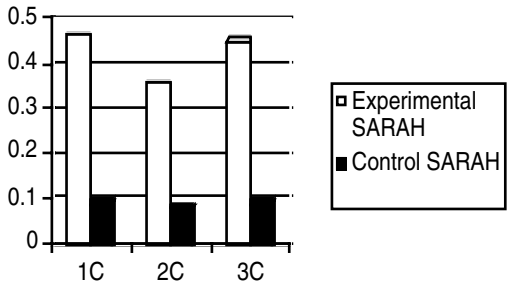


Figure 2: Simulation results for SARAH

in the Control group, $F(1,28)=37.33$, $p<.001$, ($M_e=.36$, $\pm SE_e=.04$ vs. $M_c=.09$, $\pm SE_c=.02$). The results are displayed in Figure 2. Interestingly, when running 30 SRNs with the same parameters on this task, learning of the trained sequences can be obtained, $F(1,28)=12.42$, $p<.01$, $f=.67$, $\eta^2=.31$, $M_e=.11$, $\pm SE_e=.03$ vs. $M_c=-.01$, $\pm SE_c=.01$, but no generalization to novel sequences, $F(1,28)=.44$, $p>.5$. As a result, SARAH may be better able to model the interplay between associative and cognitive processes found in (2001a).

References

Spiegel, R. & McLaren, IPL. (2001a) Human Sequence Learning: Can Associations Explain Everything? In K. Stenning & J. Moore (Eds.): *Proceedings of the Twenty-Third Cognitive Science Conference.* Mahwah, N.J.: Erlbaum.

Spiegel, R. & McLaren, IPL. (2001b) SARAH: A Cognitive Neural Model of Sequence Learning. In *Proceedings of the 5th International Conference on Cognitive and Neural Systems*. Boston University.

Spiegel, R. & McLaren, IPL. (2001c) A Hybrid Model Approach to Generalization in Sequence Learning. In *Proceedings of the International Joint INNS/IEEE Conference on Neural Networks*. Washington, D.C.

The Relationship between Learned Categories and Structural Alignment

Daisuke Tanaka (daisuke@srt.L.u-tokyo.ac.jp)
Department of Psychology, University of Tokyo; 7-3-1 Hongo Bunkyo-ku
Tokyo, 113-0033 Japan

Recent researches suggest that similarity is well characterized as a comparison of structured representations and two kinds of differences yielded through the alignment process were influenced on similarity judgement differently (Markman, & Gentner, 1996). This study applied structural alignment view to category learning and tested the hypothesis that features of categories with alignability between categories are more important than features without alignability in classification of exemplars.

Method

Subjects

18 university students participated in the experiment.

Materials & Procedure

Subjects learned a pair of categories in the learning phase. Category structure composed of short descriptions as features (Table 1). Those features could classified into 3 groups; alignable features (AF), non-alignable features (NF), and common features (CF). AF had a relation to other features composed alternative category as alignable differences. NF did not make alignable differences and were characteristic of one category. CF are in common with two categories. In the learning phase, learning exemplars were used and one learning exemplar had 3 features; one of AF, one of NF, and one of CF. Subjects were presented with the exemplars one at a time and identified them as being in category 1 or 2. After each choice, subjects were given feedback. This procedure was repeated in blocks of 18 exemplars until the subjects had correctly classified over 90% of 18 exemplars. After reaching criterion, subjects entered the test phase which was similar to the learning phase without feedback. In the test phase, test exemplars were used, which composed of "appropriate" and "inappropriate" exemplars. "appropriate" exemplars, used as fillers, could be classified one category using the knowledge of category structure, like learning exemplars. On the other hand, "inappropriate" exemplars could not be classified correctly, and divided into 3 subtypes, subtype A, subtype B, and subtypes C by the difference of component patterns of features (see Table 2).

Table 1: A part of category structure.

Category 1		Category 2
Summer sports	(AF)	Winter sports
In a group		By oneself
Indoor sports	(NF)	Popular with kids
In fashion		With ease
Need the special education		(CF)

Table 2: Examples of "inappropriate" exemplars

Subtype A	Subtype B	Subtype C
Summer sports	Summer sports	By oneself
By oneself	With ease	In fashion
Need the special education	Need the special education	Need the special education

Results and Discussion

The main result are presented in figure 1. The hypothesis of this study predicted that the subtype A exemplars were classified as members of category 1 or 2 by chance, the subtype B tended to be classified as category 1, and the subtype C as category 2. The choice tendency for category 1 was different among subtypes significantly ($F(2,34)=6.56$, $p<.01$). The percentage to be classified into category 1 in subtype B was higher than in subtype C. This result suggests that alignable features were used for two categories learning and classified exemplars into categories.

References

Markman, A. B., & Gentner, D. (1996). Commonalities and differences in similarity comparisons. Memory & Cognition, 24, 235-249.

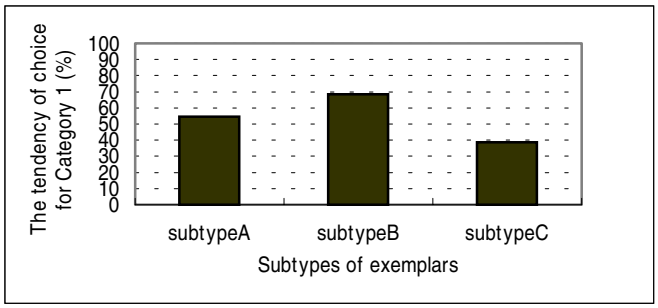


Figure 1: The choice tendency of each subtypes.

Timing and Rhythm in Multimodal Communication for Conversational Agents

Ipke Wachsmuth (ipke@techfak.uni-bielefeld.de)
Faculty of Technology, University of Bielefeld
D-33594 Bielefeld, Germany

Motivation

Synthesis of lifelike gesture is finding growing attention in human-computer interaction. In particular, synchronization of synthetic gestures with speech output is one of the goals for embodied conversational agents which have become a new paradigm for the study of gesture and for human-computer interface (Cassell et al., 2000). Embodied conversational agents are computer-generated characters that resemble similar properties as humans in verbal and nonverbal face-to-face conversation.

Gesture production in humans is a complex process leading to characteristic shape and dynamic properties of gestures which enable humans to distinguish them from subsidiary movement and recognize them as meaningful. In coverbal gestures the stroke (the most effortful part of the gesture) is tightly coupled to accompanying speech, yielding semantic, pragmatic, and temporal synchrony between the two modalities (McNeill, 1992).

Although promising work exists for the production of synthetic gestures, natural timing for the gesture stroke and synchronizing it with speech output remains a research challenge. For instance, the REA system by Cassell and coworkers (in Cassell et al., 2000) implements an embodied agent that produces verbal and gestural output. Yet though precise timing of spoken and gestural utterances is targetted in their work, the authors state that it has not been satisfactorily solved.

Articulated Communicator

A mid-range goal of our research is the conception of an "articulated communicator" that conducts multimodal dialog with a human partner in cooperating on a model airplane construction task. In this context an operational model was developed that enables lifelike gesture animations to be rendered in real time from representations of spatiotemporal gesture knowledge (Kopp & Wachsmuth, 2000). Based on various findings on the production of human gesture, the model provides means for motion representation, planning, and control to drive the kinematic skeleton of a figure which comprises 43 degrees of freedom in 29 joints for the main body and 20 DOF for each hand (see Figure 1). A movement plan is formed as a tree representation of a temporally ordered set of movement constraints in three steps:

(1) retrieve feature-based specification from a gestuary
(2) adapt it to the individual gesture context
(3) qualify temporal movement constraints in accordance with external timing constraints.

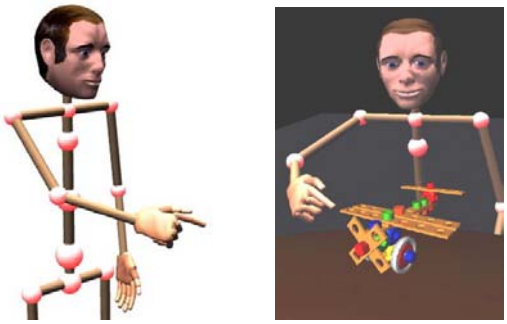


Figure 1: Articulated Communicator.

Hence our model is conceived to enable cross-modal synchrony with respect to the coordination of gestures with the signal generated by a text-to-speech system. In multimodal communication, by which we mean the concurrent formation of utterances that include gesture and speech, a rhythmic alternation of phases of tension and relaxation can be observed. The issue of rhythm in communication has been addressed widely and has been a key idea in our earlier work on synchronizing gesture and speech in HCI input devices (Wachsmuth, 1999). Achieving precise timing for accented parts in the gesture stroke as a basis to synchronize them with stressed syllables in speech is work currently in progress.

Acknowledgment

This research is partially supported by the Deutsche Forschungsgemeinschaft in the Collaborative Research Center "Situated Artificial Communicators" (SFB 360).

References

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. (Eds.) (2000). *Embodied Conversational Agents.* Cambridge (MA): The MIT Press.

Kopp, S., & Wachsmuth, I. (2000). A knowledge-based approach for lifelike gesture animation. *ECAI 2000 Proc. 14th European Conf. on Artificial Intelligence.* Amsterdam: IOS Press.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought.* Chicago: University of Chicago Press.

Wachsmuth, I. (1999). Communicative Rhythm in Gesture and Speech. In A. Braffort et al. (Eds.), *Gesture-based Communication in Human-Computer Interaction.* Berlin: Springer (LNAI 1739).

Training Task-Switching Skill in Adults with Attention-Deficit/Hyperactivity Disorder

Holly A. White (hawhite@memphis.edu)
Department of Psychology
University of Memphis
Memphis, Tennessee 38152-3230

Priti Shah (priti@umich.edu)
Department of Psychology
525 East University
University of Michigan
Ann Arbor, MI 48109-1109

Background

Task switching involves rapidly switching back and forth between multiple tasks and is part of an executive control system (e.g., Rogers & Monsell, 1995). Using a variable training approach, we found training on attention switching tasks and transfer of training to related tasks. These findings have potential for those with impairments of executive function, such as older adults or people with Attention Deficit Hyperactivity Disorder (ADHD). Kramer, Hahn, and Gopher (1999) trained younger and older adults on a task of attention switching. Results showed that age-related switch costs evident early in practice disappeared after training.

Cepeda, Cepeda, and Kramer (2000) studied task switching with ADHD and non-ADHD children and found that non-medicated ADHD children showed larger switch costs compared to non-ADHD children. Dowsett and Livesey (2000) trained children on several tasks of executive control and found that a variable training procedure resulted in improvement of generalized response capabilities, such as inhibitory control. The present study applied the variable training procedure to adults with characteristics of ADHD and demonstrated that training effects are not limited to task-specific strategies.

Methods

Undergraduate students from the University of Memphis were recruited to participate in the study on the basis of their responses to a questionnaire used to assess adult ADHD. Participants were divided into ADHD and non-ADHD groups and assigned to either the training or control condition. The training condition included a pretest, six training blocks, and a posttest. The control condition included a pretest, six filler blocks (non-executive control tasks), and a posttest. The training tasks were variants of the number-letter task used by Rogers and Monsell (1995).

Results and Discussion

The ADHD and non-ADHD groups showed training effects greater than the control groups. Additionally, training effects transferred to related tasks of attention switching. These results may have implications for efforts directed at cognitive rehabilitation of ADHD individuals. In future research, we would like to examine the mechanisms underlying training, perhaps using neuroimaging. Such research may help to determine whether subjects are using compensatory strategies or actually improving some underlying attentional process. Future research will also be directed at examing the scope of training. Specifically, how much transfer occurs for less related tasks (e.g., WCST)? How long do the effects of training last, and will children show training comparable to that seen in adults? Perhaps training, in the form of cognitive rehabilitation, will be helpful in place of, or in combination with, the stimulant medications traditionally used to treat ADHD.

References

Cepeda, N.J., Cepeda, M.L., & Kramer, A.F. (2000) Task switching and attention deficit hyperactivity disorder. *Journal of Abnormal Child Psychology, 28,* 213-226.

Dowsett, S.M., & Livesey, D.J. (2000). The development of inhibitory control in preschool children: Effects of "executive skills" training. *Developmental Psychobiology, 36,* 161-174.

Kramer, A.F., Hahn, S., & Gopher, D. (1999). Task coordination and aging: Explorations of executive control processes in the task switching paradigm. *Acta Psychologica, 101,* 339-378.

Rogers, R.D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General, 124,* 207-231.

Advantages of a Visual Representation for Computer Programming

Kirsten N. Whitley (whitley@vuse.vanderbilt.edu)
Department of Computer Science, Box 1679 Station B
Nashville, TN 37235 USA

Laura R. Novick (Laura.Novick@vanderbilt.edu)
Department of Psychology, Box 512 Peabody
Nashville, TN 37203 USA

Doug Fisher (dfisher@vuse.vanderbilt.edu)
Department of Computer Science, Box 1679 Station B
Nashville, TN 37235 USA

Pictures and diagrams have long played an important role in human societies (Novick, 2001). Prehistoric peoples painted pictures on cave walls. The Bayeux Tapestry, from the 11th century, records the events surrounding the Battle of Hastings. In the 15th century, da Vinci made thousands of anatomical, mechanical, geographical, and other drawings. Today, diagrams are found on blackboards in most university departments (McKim, 1980). In our increasingly technical and technological society, diagrams (especially abstract ones) are likely to gain in importance. Even the traditionally text-driven field of computer programming now includes languages whose representations are diagrammatic rather than textual (Whitley, 1997).

LabVIEW™ (National Instruments, 1998), for example, features a programming language based on the dataflow paradigm in which the flow of information through the program is expressed using a notation that resembles circuit diagrams. LabVIEW™ was designed to facilitate the development of data acquisition, analysis, display, and control applications for science and engineering laboratories.

In an experiment with a 2 × 3 mixed-factors design, we assessed the comprehensibility of LabVIEW™'s representation. Representation type was manipulated between subjects, with 15 upper-level computer science students randomly assigned to receive LabVIEW™'s visual representation and 16 such students randomly assigned to receive an equivalent textual dataflow representation. The second factor – problem type (tracing, parallelism, debugging) – was manipulated within subjects.

The experiment involved a 90 min lecture, during which subjects were taught a subset of the LabVIEW™ language, followed by a 90 min test session. The test problems required subjects to read and understand code segments. For the three tracing problems, subjects were given input values for variables in the code and had to determine what the output values would be if the code were to execute. For the three parallelism problems, several program operators were highlighted, and subjects had to specify the order in which pairs of operators could execute. For the three debugging problems, subjects were given written specifications for the code and had to find the (single) error in the code.

For both solution accuracy and time, representation type interacted with problem type. For the more difficult parallelism and debugging problems, the visual representation was clearly superior to the textual representation: The visual subjects had higher accuracy scores and spent less time working on the problems. For the tracing problems, accuracy was similar for the two representations, but the visual subjects spent more time working on them. Overall, these results provide compelling evidence for the superiority of LabVIEW™'s visual representation over an equivalent textual representation.

The comprehensibility effects we found seem likely to generalize to novice LabVIEW™ programmers in more naturalistic situations. The subset of LabVIEW™ we used included a fair portion of the language. Moreover, our experimental tasks were representative of actual programming tasks and were sequenced in a natural instructional order. Although it is an open question whether the superiority of the visual representation will apply to other types of programming tasks (e.g., writing new code, modifying existing code), we suspect that it will under a variety of conditions.

References

McKim, R. H. (1980). *Thinking visually: A strategy manual for problem solving*. Belmont, CA: Wadsworth.

National Instruments (1998). LabVIEW™ User Manual [Software manual]. Austin, TX: Author.

Novick, L. R. (2001). Spatial diagrams: Key instruments in the toolbox for thought. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 40, pp. 279-325). San Diego, CA: Academic Press.

Whitley, K. N. (1997). Visual programming languages and the empirical evidence for and against. *Journal of Visual Languages and Computing, 8,* 109-142.

Mass and *Count* in Language and Cognition: Some Evidence from Language Comprehension

Heike Wiese (heike.wiese@rz.hu-berlin.de)
Humboldt University Berlin, Department of German Language and Linguistics, Unter den Linden 6
100999 Berlin, Germany

Maria M. Piñango (maria.pinango@yale.edu)
Yale University, Department of Linguistics, P.O.Box 208236, HGS 318
New Haven, CT, USA

In linguistics and the philosophy of language, the mass/count distinction has traditionally been regarded as a bi-partition on the nominal domain, where typical instances are nouns like 'beef' (*mass*) vs. 'cow' (*count*).

In the present paper, we argue that this partition reveals a system that is based on both syntactic features and conceptual features, and present experimental evidence suggesting that the discrimination of the two kinds of features has a psychological reality.

We account for the mass/count distinction by a binary classification of nouns based on a syntactic feature [± tn] ('transnumeral') and a conceptual feature [± mn] ('mass'), with the following diagnostics: Nouns are [-tn] if and only if they obligatorily occur in their plural form when denoting more than one realisation of the nominal concept. Nouns are [+mn] if they refer to homogeneous masses, and [-mn] if they refer to objects.

According to this classification, 'beef' is [+tn, +mn] and 'cow' is [-tn, -mn]; syntactic differences go together with conceptual differences here. However, this is not necessarily so. Collective nouns like 'cattle' behave like 'beef' syntactically, but fall into one class with 'cow' conceptually: they are not marked for number, and are therefore [+tn], but they refer to objects, i.e., they are [-mn]. Hence conceptual features are not in a one-to-one correspondence with syntactic features:

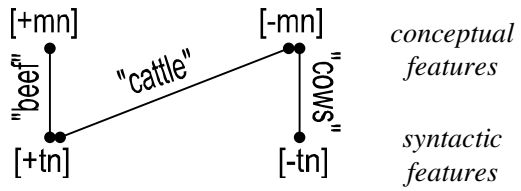


Figure 1: Dissociation of conceptual and syntactic correlates of the mass/count distinction

Does this distinction have a psychological reality? We addressed this question by investigating whether facilitation of lexical activation (in the form of priming) can be obtained for [-mn] versus [+mn] nouns in language comprehension: we investigated whether exposure to a [-mn] noun (the *prime*; e.g. 'furniture') reduces the time needed for the subsequent activation of another [-mn] noun (the *related target*; e.g. 'cattle'), in comparison to the activation of a [+mn] noun (the *unrelated target*; e.g. 'beef').

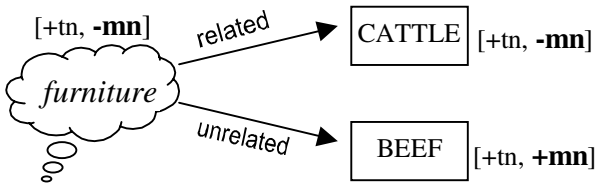


Figure 2: Experimental prime-target pairs

Since the [-mn] nouns we used were collectives (i.e., they were syntactically [+tn], like all [+mn] nouns), the difference between related and unrelated targets was restricted to the conceptual feature [± mn].

We conducted two studies. In study 1, primes were presented auditorily (via headphones) in a sentential context; in study 2, primes were presented visually (on a computer screen) as isolated words. All targets were presented visually and appeared immediately after the prime was heard (study 1) or seen (study 2). Subjects performed a lexical decision on the targets (and on non-experimental probes), i.e., they had to decide whether they saw a word or a non-word. Reaction times were measured for related versus unrelated targets.

In both studies, after [-mn] primes reaction times were faster for [-mn] targets than for [+mn] targets: facilitation for [± mn] was evident both in sentential contexts (study 1) and in word lists (study 2), and for auditory input (study 1) as well as for visual input (study 2).

We interpret this as evidence that the [± mn] distinction, as a conceptual correlate of the nominal mass / count partition, has a psychological reality independently of the syntactic distinction of nouns.

Acknowledgements

The research presented here was supported by a TransCoop grant awarded to the authors by Alexander von Humboldt Foundation, by a DAAD postdoctoral grant to Heike Wiese, and by NIH Grant DC 03660 to Braindeis University.

Inhibition Mechanism of Phonological Short-term Memory in Foreign Language Processing

Takashi Yagyu (yagyu@srt.L.u-tokyo.ac.jp)
Department of Psychology, University of Tokyo; 7-3-1 Hongo Bunkyo-ku
Tokyo, 113-0033 Japan

This study examined phonological short-term memory (phonological STM) capacity and the individual difference of the range for inhibition. About the relation between phonological STM capacity and foreign language processing and acquisition, there are many antecedence studies (e.g., Baddeley, Gathercole & Papagno, 1998). However, the degree of similar of a native language and a foreign language and the relation of phonological STM are seldom clarified. Then, in this research, it investigated the relation between the degree of similar of the item to memorize and an interference item, and the individual difference of inhibition capability.

Method

Subjects

The subjects were 25 undergraduate and graduate students. They were all native Japanese speakers who had studied English for 6-10 years.

Materials

<Word-Span Task>The phonological STM capacity was measured with the use of a word-span task written in Japanese and English. In each span task, the first three sets consisted of three words, the next three sets of four words etc. The largest set size was seven words, making total number of items 75.

<Dual-Task of Word Retention> A total of 144 word-sentence pairs were grouped into four lists of 36 pairs each. The first list contained Japanese words and English sentences, and second list contained English words and Japanese sentences. The other two lists were same language structure in word-sentence pairs. In each list, the first four sets consisted of two word-sentence pairs, the next four sets of three word-sentence pairs etc. The largest set size was four word-sentence pairs.

Table 1: Correlation with the results of a dual-task and of the word-span task(n=25).

dual-task (word-sentence)	L1-L1	L1-L2	L2-L1	L2-L2
WST	0.56**	0.37+	0.30 n.s.	0.65**

+ p < .10 **p < .01

Procedure

<Word-Span Task>Word span was assessed in both Japanese and English. On each measure, subjects were required to recall a sequence of words immediately after visual presentation.

<Dual-Task of Word Retention>A primary task was to recall words. Reading sentences aloud have been used as a method of the background task. First words, first sentence, second word, second sentence were presented successively. After reading all words and sentences, the subject tries to recall.

Results and Discussion

Correlations between the Japanese and English version word-span tasks showed significant at $p<.01$ level($r=.51$). It divided into the high-span group and the low-span group by the mean score of the results of the Japanese and English version word-span task.

The main results are presented Figure 1. It is that the difference was found by the results pattern of a retention word between the group with high-span , and the low-span group with the combination of the language kind of a retention word and interference sentence. Then, the correlation with the results of a dual-task and the results of the word-span task was investigated (Table 1). It was suggested the memorizable number of words not only increases that a phonological STM capacity is large, but that it can be inhibited to what has the high degree of similar of the information used as interference.

References

Baddeley, A. D., Gathercole, S. E., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, **105**, 158-173.

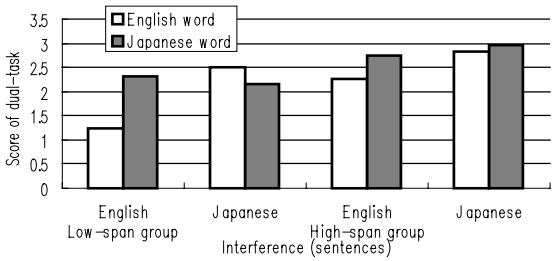


Figure 1: Score of the dual-task between each condition .

Odd-Even effect in multiplication revisited: The role of equation presentation format

Michael C. W. YIP

School of Arts & Social Sciences, The Open University of Hong Kong

myip@ouhk.edu.hk

Introduction

Several studies suggested that people usually make use of many different strategies to solve simple arithmetic problems (digit-addition and digit-multiplication). (Lemaire & Fayol, 1995; LeFevre, Bisanz, Daley, Buffone, Greenham, & Sadesky, 1996; Krueger, 1986; Krueger & Hallford, 1984; Zbrodoff & Logan, 1990). For example, to verify simple multiplication questions, 2 x 3 = 7; 4 x 6 = 25, people will retrieve the answer of the question from their memory base and compared it with the mentioned answer (memory-retrieval hypothesis) *or* people will first check out if the mentioned answer of the question violates the parity rules of multiplication (parity-checking hypothesis). Recently, a similar study was conducted by Lochy, Seron, Delazer, & Butterworth, (2000), they proposed an alternative account to the odd-even effect: number evenness hypothesis. They reasoned the effects were mainly due to the familiarity of even numbers of the products rather than the parity-checking rule. However, the rationale of this hypothesis is basically in line with the memory-retrieve hypothesis because both of them emphasized the distributional information of the products or the mathematical equation from the memory. To further verify the odd-even effect in multiplication and to assess the validity of those hypotheses, I replicated the study by altering the presentation format of the mathematical equation in this paper. revising from the traditional standard format [a x b = c] to a reverse format [c = a x b]. The reverse presentation format will obviously affect the memory capacity of the equation because people would not normally remember the multiplication table in that specific format. Therefore, if the memory-retrieval hypothesis or the number evenness hypothesis are really the underlying mechanism of the multiplication processes, the similar pattern of results will be showed and should be consistent with the previous studies (Lochy et al., 2000). Otherwise, other explanations should be further sought. The present study attempts to evaluate the strengthen of those hypotheses and their roles played in the temporal course of simple multiplication process by using the presentation format as a crucial examination tool.

Experiment

The basic design of the present experiment is similar to the study of Lochy et al. (2000). Three main variables in the present experiment: (1) presentation format of the equation: (a x b = c) vs. (c = a x b); (2) types of problem: (even x even) vs. (odd x odd) vs. mixed; (3) size of split: +1 +2 +3 +4 –1 –2 –3 –4.

Procedure

A series of simple multiplication problems were randomly presented to each participant in one of the two forms (a x b = c) or (c = a x b). Participants were asked to verify whether the equation is true or false by pressing a key. Response latencies were recorded from the onset time of the equation that displayed on the computer screen to the manual response.

Results and Discussion

Three main findings in the present study were concluded.

First, the format of equation presentation is in fact influence the verification time of the participants (reverse mode takes longer time to verify than the traditional mode). It is sensible because the internal representation of the simple multiplication knowledge is initially encoded in the traditional form (2 x 3 = 6) instead of the reverse form (6 = 2 x 3).

Second, collapsed over the levels of presentation format, in the traditional format, results obtained from the present study were replicated the general pattern of results to the odd-even effects which were reported in Lochy et al., (2000). Third, the most interesting point here is that on the contrary, in the reverse presentation mode, the odd-even pattern of results was typically consistent with the parity-checking rule rather than the other hypotheses. Clearly, in the processing of simple multiplication, people will rely heavily on the mathematical knowledge stored in their memory base. However, once the relevant information cannot be easily triggered from the memory store, other alternatives will be used immediately (parity-checking rule) to solve the original arithmetic problem. These findings seem to support the multiple-strategy hypothesis in solving simple mathematical problems (Lemaire & Fayol, 1995).

References

Krueger, L. E. (1986). Why 2 × 2 = 5 looks so wrong: On the odd-even rule in product verification. *Memory & Cognition, 14*, 141-149.

Krueger, L. E., & Hallford, E. W. (1984). Why 2 + 2 = 5 looks so wrong: On the odd-even rule in sum verification. *Memory & Cognition, 12*, 171-180.

LeFevre, J., Bisanz, J., Daley, K. E., Buffone, L., Greenham, S. L., & Sadesky, G. S. (1996). Multiple routes to solution of single-digit multiplication problems. *Journal of Experimental Psychology: General, 125*, 284-306.

Lemaire, P., & Fayol, M. (1995). When plausibility judgments supersede fact retrieval: The example of the odd-even effect on product verification. *Memory & Cognition, 23*, 34-48.

Lochy, A., Seron, X., Delazer, M., & Butterworth, B. (2000). The odd-even effect in multiplication: Parity rule or familiarity with even numbers? *Memory & Cognition, 28,* 358-365.

Zbrodoff, N. J., & Logan, G. D. (1990). On the relation between production and verification tasks in the psychology of simple arithmetic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 83-97